# Methods and Data Analysis 3

## Introduction

It is known fact today that smoking by mothers can lead to a variety of health issues in babies. One of the first studies conducted to analyze this claim was done fifty years ago at Kaiser Foundation Hospital in Oakland, CA. This study collected data over a span of seven years from familes with different socio economic backgrounds. The goal of this assignment is to analyze what parameters play a imporatant role in the overall health of the babies and importantly premature births. We also examine if mother's smoking habits have any particular effect on baby's health.

## Data:

The original data provided has information columns about mother as well father. But the father information columns are empty for many data points. For the purpose of this study we have ignored those columns. The data used in this assignment has 12 fields. The data dictionary is as follows:

1. **id:** Index Number.
2. **date:** Birth date (number of days from January1, 1961) *(discrete variable)*
3. **gestation:** Length of gestation in days. *(discrete variable)*
4. **bwt.oz:** Birth weight in ounces. *(continuous variable)*
5. **parity:** Total number of previous pregnancies, including fetal deaths and still births. *(discrete variable)*
6. **mrace:** Mother's race *(categorival variable)*
7. **mage:** Mother's age *(discrete variable)*
8. **med:** Mother's education *(discrete variable)*
9. **mht:** Mother's height *(discrete variable)*
10. **mpregwt:** Mother's pre-pregnancy weight in pounds *(discrete variable)*
11. **inc:** Family income *(categorical variable)*
12. **smoke:** Mother's smoking habit *(categorical variable)*
13. **premature:** Baby is premture or not (outcome). *(discrete variable)*

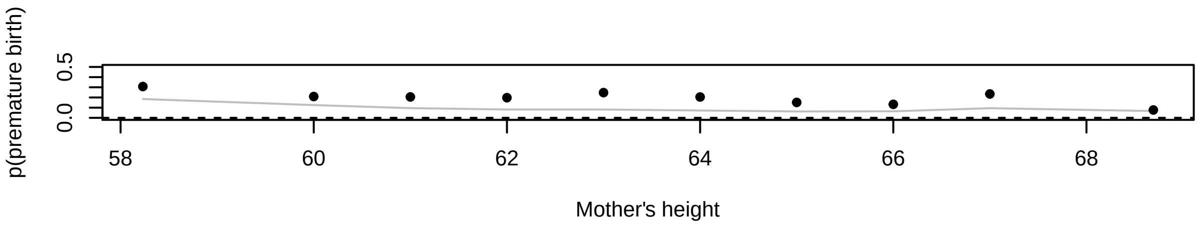
The variables *mrace*, *med*, *income* and *smoke* in the data are discrete, which have been converted to categorical variables as they do not have a relative order.

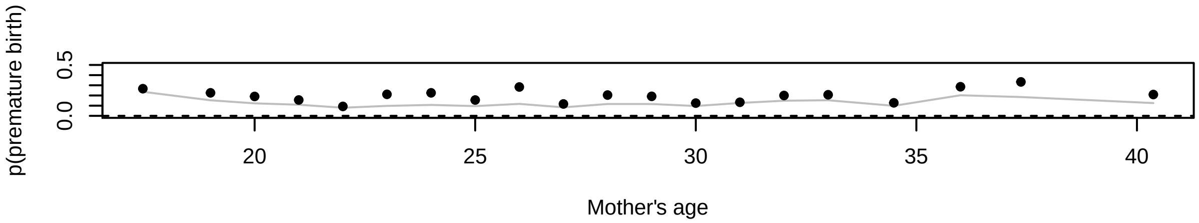
In this assignment one of the important questions we want to answer is if smoking has any effect on premature births. We use contingency table to see conditional probabilities and see that conditional probability of premature birth given smoking mothers is 0.21 and conditional probability of premature birth given non-smoking mothers is 0.16. The difference is not very high, so we further evaluate the relationship between variable *smoke* and *premature* using Chi-square test.

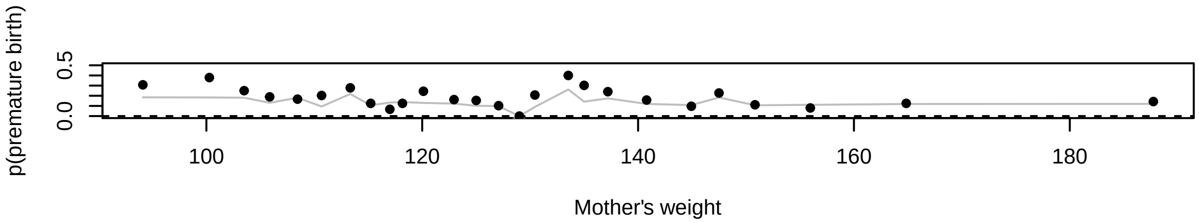
To analyze the association between the categorical variables and the outcome variable *premature*, Chi-square tests were performed for these variables. The following table summarizes the results of the Chi-squared test. From the table we see that predictors *mrace*,*med* and *parity* have association with *premature* variable, whereas *inc* and *smoke* has a high p-value and hence is independant from *premature*. However since we need to find the effect of smoking on prematuer births we include it in our analysis. Hence in our basic model we will exclude *inc* but will include other cateogorical variables.

| variable | p.value | Significant |
|---|---:|---|
| smoke | 0.0694000 | No |
| mrace | 0.0035610 | Yes |
| med | 0.0005476 | Yes |
| parity | 0.0112500 | Yes |
| inc | 0.9087000 | No |

For continuous variables, binned plots were plotted to check if any transformations are required, and how the probabilitity of being mature varies with the continuous predictors. For the variables *mage* and *mht* the binned plot has almost linear trend, however for *mpregwt* the binned plot shows high variations. This might be due to very few data points in higher valued bins.







To check for interactions between the categorical variable *smoke* and continuous variable *mage*, *mht* and *mpregwt*, nothing significant difference can be seen for *mage* and *mpregwt*. For *mht* visually there seems to be some interaction with *smoke* (plots in Appendix) We will evaluate if this interaction term is significant in the model.

We will mean center the continuous variables so that the weights of predictors can be interpreted in a meaningful way.

# Model

1. Model 1: First we build a model using all the variables. In the next step we AIC method to remove variables which are not significant.

```
modelbasic <- glm(premature ~ mpregwt_c + mht_c + parity + mage_c + med + inc + smoke + mrace, data = sm
Conf_mat_basic <- confusionMatrix(as.factor(ifelse(fitted(modelbasic) >= mean(smoking$premature), "1","(
```

2. Model 2: With the above model using all the variable as the full model we use the AIC method in step-wise fashion to find the optimal predictors. The AIC method tries removing and adding the variables back and forth to obtain the optimal predictors. Using this method the significant predictors are *mrace*, *mpregwt*, *med*, *smoke*.
3. Model 3 Next we try to add the interacion between *smoke* and *mht* and see if the interaction term is significant by comparing with model 1 using anova test. The p-value of anova test is high and hence we conclude that the interaction term is not significant and should not be included in our model.
4. Model 4 We also try the intercation term for *smoke* and *mrace* as we want to check if this association has any significance.

The following table compares the accuracy, sensitivity, specificity and area under the ROC curve for the four models created above.

| ModelName | Accuracy | Sensitivity... | Specificity... | AUC... |
|---|---|---|---|---|
| Model 1 | 60.07% | 57.93% | 60.57% | 63.8% |
| Model 2 | 60.03% | 59.76% | 60.43% | 63.1% |
| Model 3 | 63.29% | 55.49% | 60.85% | 63.8% |
| Model 4 | 59.15% | 59.76% | 59.01% | 63.8% |

In this problem comparing just accuracy to select the best model might not be a correct idea. Sensitivity seems to be of importance as identifying premature births is important. And since the premature variable is skewed (few points for premature=1) increase in sensitivity reduces overall accuracy of the model. Using the above value comparisons, model 2 (model obtained using AIC) is selected as the final model. Predictors used in this model are *med*, *mpregwt*, *smoke*, *mrace*.

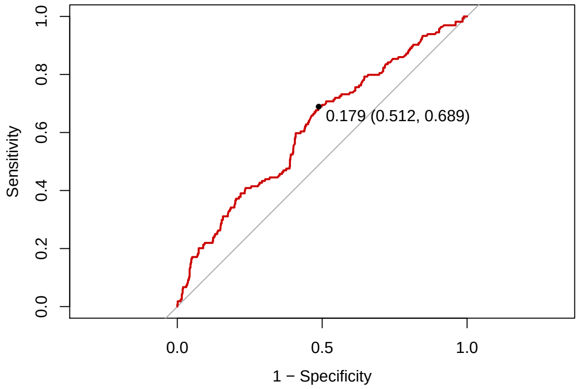
Next, the values of the weight estimates, standard error, z value and p-value are summarized in the table below for model 2.

| term | estimate | std.error | statistic | p.value |
|---|---:|---:|---:|---:|
| (Intercept) | -1.5681331 | 0.2564632 | -6.114457 | 0.0000000 |
| mrace | 0.0865358 | 0.0283139 | 3.056300 | 0.0022409 |
| mpregwt_c | -0.0105177 | 0.0044988 | -2.337882 | 0.0193934 |
| med | -0.1219881 | 0.0633060 | -1.926959 | 0.0539848 |
| smoke | 0.3311379 | 0.1796060 | 1.843691 | 0.0652281 |

The ROC curve for the final model can be found below:

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

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



For the final model we analyze the binned plots to check for any patterns in the residuals. (plots in appendix) 1. From the residuals vs. fitted values there is no evident pattern in the plot, however three points lie outside the confidence interval and one point lies on the interval. 2. The residuals vs. mpregwt plot does not have any specific pattern that we should have accounted for. Two points lie outside the confidence interval. 3. The residuals vs. med too does not have any evident pattern. 4. The mean residuals for *mrace* also do not have any specific pattern.

## Conclusion:

Question1: *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?*

Answer: In the final model the coefficient of variable smoke is approximately 0.31 which implies the increase in odds ratio of premature birth for mothers who smoke vs. who do not smoke is 36% (keeping *med*, *mrace* and *mpregwt* constant). 95% confidence range for the odds ratio of pre-term birth for smokers is (0.95, 1.94)

Question2: *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.*

Answer: In model 4 we added the interaction term between *smoke* and *mrace*. But the p-values for these interaction terms is high and hence the interaction is not statistically significant. This is probably because the number of data points are very high in "white" category and low for others. Due to such unequal distribution of data we have not used it in the final model

Question3: *Are there other interesting associations with the odds of pre-term birth that are worth mentioning?*

Answer: Apart from *smoke* predictor, *med*, *mrace* and *mpregwt* have association with *premature*. For every unit increase in mother's education level the odd's ratio decreases by 12.6% (*mpregwt*, *mrace*, *smoke* held constant). Whereas, for every unit increase in mother's weight odd's ratio decrease by 1.1%. For mother's race (baseline being "white" race):

1. Comparing "black" with baseline race, odd's ratio increases by 103% (other variables held constant)
2. Comparing "asian" with baseline race, odd's ratio increases by 156% (other variables held constant)

Races "mexican" and "mix" are not statistically significant and hence their interpretation is not mentioned.

# Limitation:

1. Since the premature column was skewed (18% data points for premature=1) and our primary metric not being accuracy, the overall accuracy of the final model is not very high. We have tried to maintain both sensitivity and specificity as high as possible together and due to this total accuracy is low.

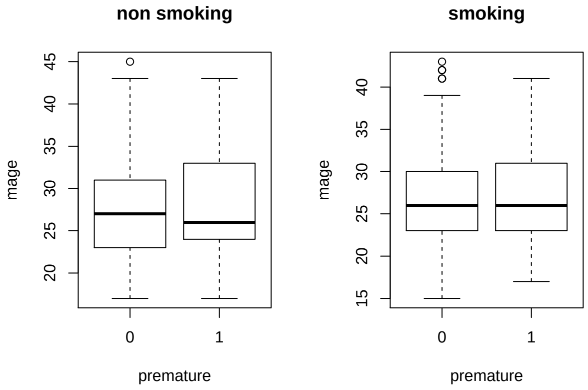
2. The number of data points(as shown in table below) are very high in "white" category and low for others.

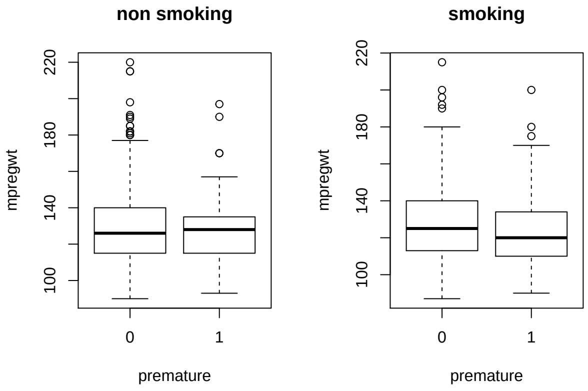
```
##
##        0   1   2   3   4   5   6   7   8   9
##   0 327  31  13  40  33  81  19 124  23  14
##   1  62   3   5   4  11  16   6  45  11   1
```

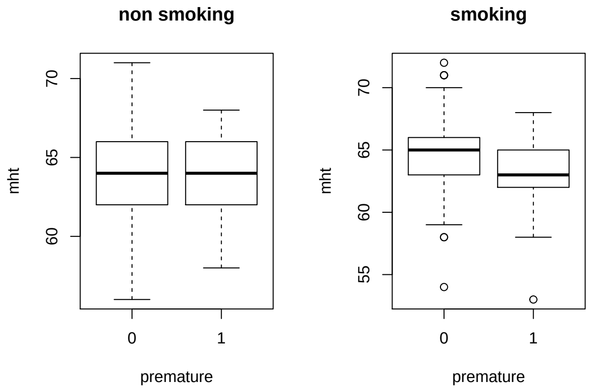
As a result, interaction term of *smoke* and *mrace* was not reliable

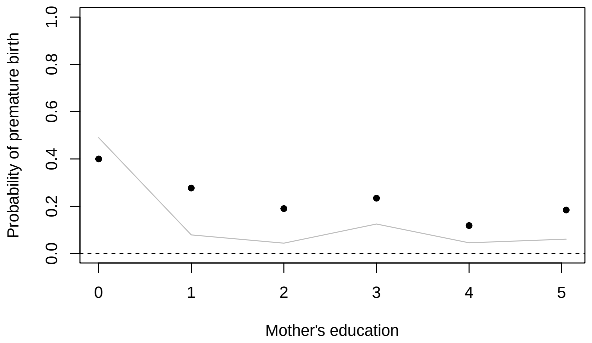
3. The *mpregwt* column also has enequal distribution within the total range of column values.

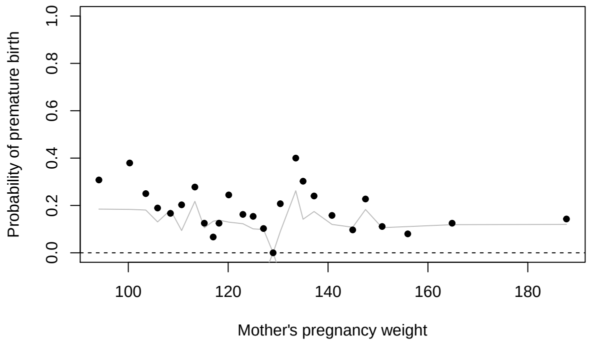
# Appendix



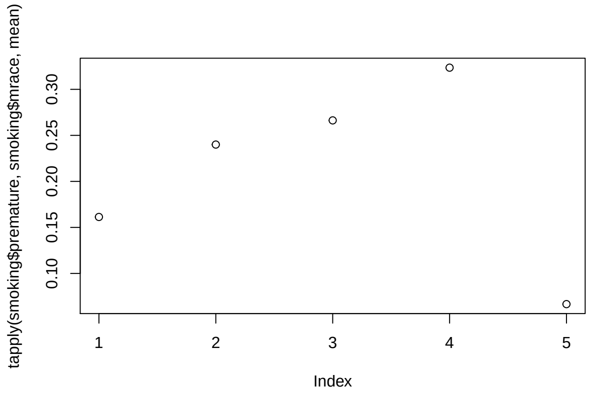


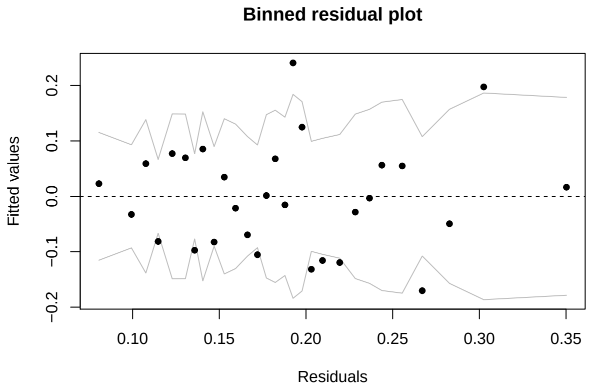


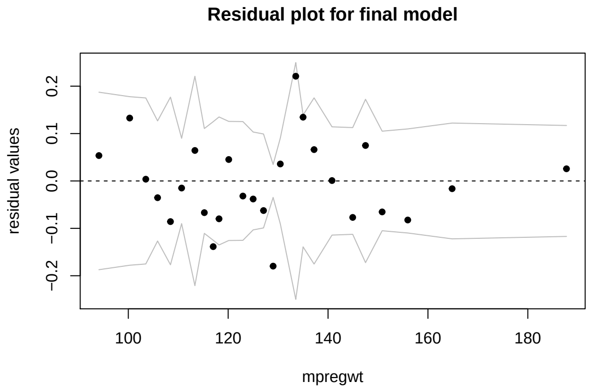


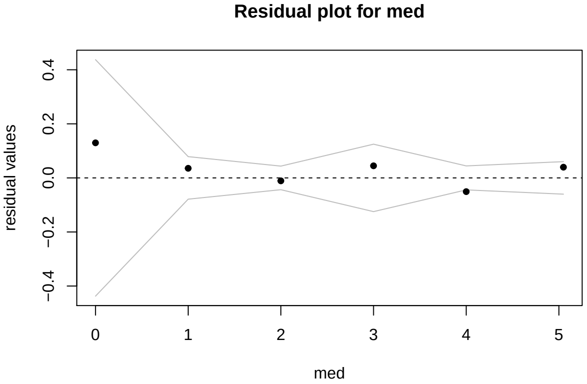


```
##           mrace
## premature white mexican black asian mix
##         0   525      19   124    23  14
##         1   101       6    45    11   1
```

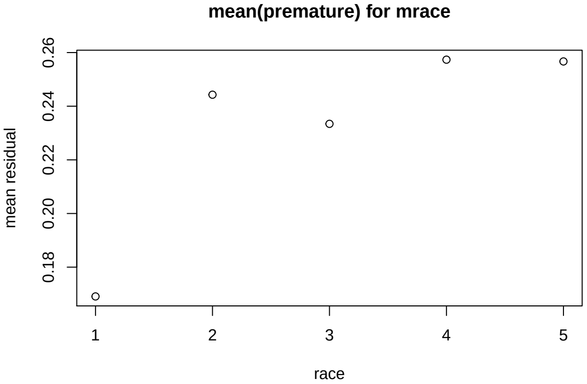



Binned residual plot


Residual plot for final model



```
##         0         1
## 0.1652361 0.2158809
```


mean(premature) for mrace
mean residual
race