Methods and Data Analysis 3

Sebastián Soriano Pérez [ss1072]

9/27/2019

MATERNAL SMOKING AND PRE-TERM BIRTH

- **Summary**

By analyzing the data on 869 newborn male babies and their families, a model was created with stepwise selection using BIC as a comparison parameter to interpret and associate the variables that were found to be significant with the response variable of a birth being premature (< 270 days of gestation). Afterwards, the model's accuracy, sensitivity and specificity were compared to a model including the variable mht. The new model improved these values marginally, so it was selected for the data analysis.

The final model estimates that only the variable of mracewhite is significant, but the rest of the mrace variables as well as med, mpregwt_c, smoke, and mht were included because they improve the model overall. The specific coefficient values can be found in the "Model" section.

- **Introduction**

This document presents a model to interpret the impact of several variables on a newborn's chances of being premature. A dataset was analyzed considering the available data in order to find the best model to explain the association between the predictive variables and the response variable through an initial exploratory data analysis (EDA), and later with a stepwise selection in R a logaritmic regression to estimate the probability of being premature. The main focus of this document is to find whether or not smoking during pregnancy had an impact in the chances of having a pre-term birth, and if this chances differ by race.

- **Data**

The Child Health and Development Studies research was one of the first to collect data to understand and quantify the risk of smoking during pregnancy to the baby's health. The data was collected from 1960 to 1967, and a subset of that data is being analyzed in this document (the variables related to the father's information are neglected for this analysis). 869 cases of newborn male babies who lived at least 28 days are being analyzed (data set smoking.csv). The purpose of this document is to present a statistical model to interpret and understand the correlation between several variables and the chances of having a pre-term birth (< 270 days). The variables being considered for building the model, in association to the response variable for a logarithmic regression model of the probability of having a pre-term birth (premature), are the following:
- Total number of mother's previous pregnancies (parity) (numeric)
- Mother's race or ethnicity (mrace) (categorical)
- Mother's age in years at pregnancy termination (mage) (numeric)
- Mother's education level (med) (categorical)
- Mother's height in inches (mht) (numeric)
- Mother's pre-pregnancy weight in pounds (mpregwt) (numeric)
- Family yearly income in 2500-increment categories (inc) (categorical)
- Indicator for the mother's smoking (smoke) (categorical)

A summary of the data variables being analyzed can be found in Annex 1.1. An exploratory data analysis for all variables and plots for their interactions can be found in Annex 1.2.

The EDA suggests none of the numerical variables have a clear association with premature as the boxplots for premature = 0 and premature = 1 do not have noticeable differences. For the categorical varibles, there are more interesting results in the conditional probability tables for each variable and their association with premature. This suggests that the categorical variables should be included in the model to evaluate their significance. The numerical variables do not need any obvious transformations as all of them suggest linear trends. The interactions parity_c:mage_c, parity_c:mpregwt_c, mage_c:mpregwt_c, mht_c:mpregwt_care being considered as those predictors have the largest correlations as seen in Annex 1.1's correlation table.

- **Model**

Various methods for model selection were tested and interactions discussed in the Data section were considered as part of the full model. Ultimately, a stepwise selection with BIC model and a model including the mht variable (which was thought to be significant) were compared and the one with the best accuracy, sensitivity, and specificity values was selected as the final model. The following was the final model that was obtiained (for the R output, see Annex 1.0):

$$\log\left(\frac{\hat{\pi}_i}{1+\hat{\pi}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1(med_1)_i + \hat{\beta}_2(med_2)_i + \hat{\beta}_3(med_3)_i + \hat{\beta}_4(med_4)_i + \hat{\beta}_5(med_5)_i + \hat{\beta}_6(med_7)_i$$
$$+\hat{\beta}_7(mrace_{black})_i + \hat{\beta}_8(mrace_{mexican})_i + \hat{\beta}_9(mrace_{mix})_i + \hat{\beta}_{10}(mrace_{white})_i$$
$$+\hat{\beta}_{11}mpregwt_i + \hat{\beta}_{12}(smoke_1)_i + \hat{\beta}_{13}mht_i$$

Where $(med1)_i + (med2)_i + (med3)_i + (med4)_i + (med5)_i + (med7)_i \in \{0,1\}$,
$(med_1)_i \in \{0,1\}; (med_2)_i \in \{0,1\}; (med_3)_i \in \{0,1\}; (med_4)_i \in \{0,1\}; (med_5)_i \in \{0,1\}; (med_7)_i \in \{0,1\}$,
$(mrace_{black})_i + (mrace_{mexican})_i + (mrace_{mix})_i + (mrace_{white})_i \in \{0,1\}$,
$(mrace_{black})_i \in \{0,1\}; (mrace_{mexican})_i \in \{0,1\}; (mrace_{mix})_i \in \{0,1\}; (mrace_{white})_i \in \{0,1\}$,

Table 1: Table of coefficients beta_0 to beta_13

	x
(Intercept)	-0.0662501
med1	-0.4994131
med2	-0.8465865
med3	-0.6587618
med4	-1.4798254
med5	-1.0056888
med7	1.8573586
mraceblack	-0.1086510
mracemexican	-0.7514873
mracemix	-1.6438913
mracewhite	-0.8567873
mpregwt_c	-0.0104869
smoke1	0.2952453
mht_c	-0.0283785

Table 2: Table of the 95% confident intervals for the coefficients beta_0 to beta_13

	2.5 %	97.5 %
(Intercept)	-2.0835527	1.9510525

	2.5 %	97.5 %
med1	-2.3715520	1.3727258
med2	-2.7025540	1.0093809
med3	-2.6183157	1.3007920
med4	-3.3718225	0.4121718
med5	-2.9005981	0.8892204
med7	-1.0572315	4.7719487
mraceblack	-0.9930631	0.7757612
mracemexican	-1.9995199	0.4965454
mracemix	-3.8355524	0.5477697
mracewhite	-1.6674040	-0.0461705
mpregwt_c	-0.0210473	0.0000735
smoke1	-0.0665268	0.6570175
mht_c	-0.1104818	0.0537247

The only significant variable is mracewhite with a p-value of 0.0383. The other variables were included to improve the model's accuracy, sensitivity, and specificity. The AIC of the model is 823.45.

The model has the following predicting parameters: Accuracy = 0.8147296, Sensitivity = 0.02439024, Specificity = 0.99858156 with a 0.5 cutoff. With a mean cutoff, the model has the following values: Accuracy = 0.6214039, Sensitivity = 0.5975610, Specificity = 0.6269504.

Finally, the ROC curve shows a value of 0.6593. For the full R output see Annex 1.0.

- **Conclusions and Remarks**

According to this model, it cannot be concluded that mothers who smoke tend to give birth prematurely compared to mothers who do not smoke. The coefficient for smoke1 was not significant (the p-value of 0.1097 indicates we cannot reject the null hypothesis that the coefficient beta_12 is equal to zero). There is not enough evidence to suggest smoking affects the log odds (and thus the odds) of there being a premature birth (< 270 days).

Table 3: Table of coefficients beta_0 to beta_13 in the odds scale

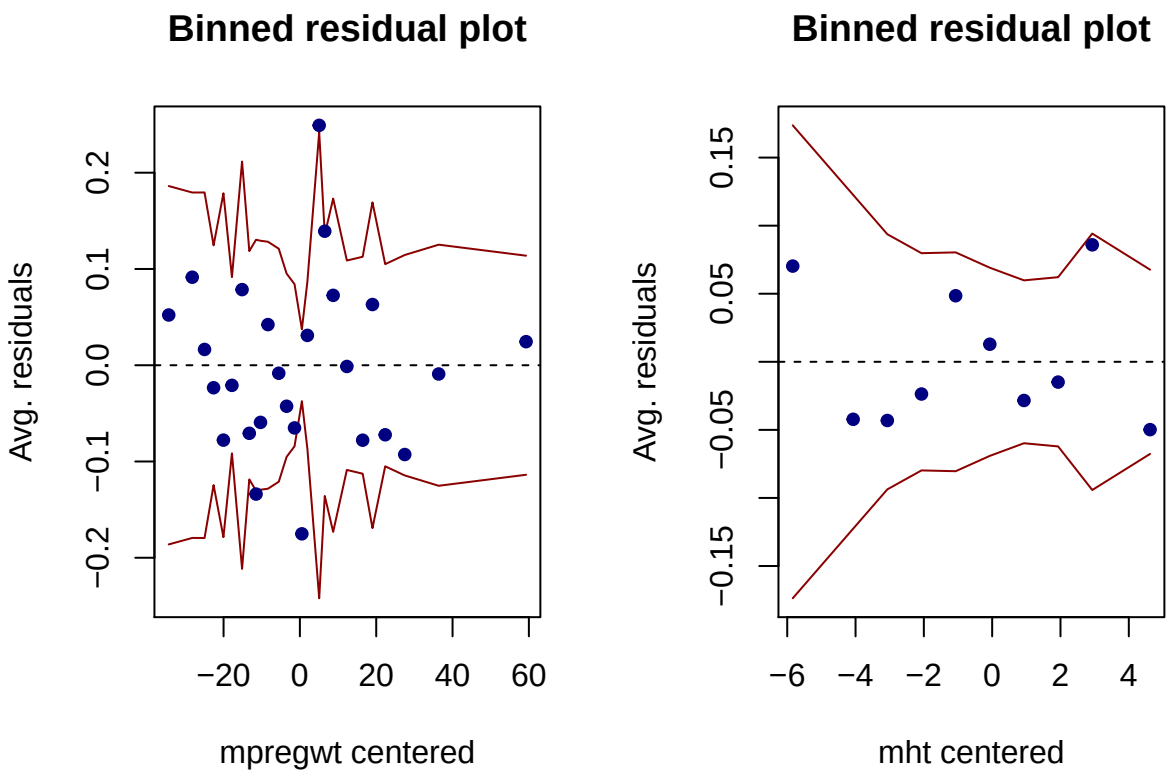
	x
(Intercept)	0.9358967
med1	0.6068867
med2	0.4288764
med3	0.5174917
med4	0.2276774
med5	0.3657926
med7	6.4067917
mraceblack	0.8970434
mracemexican	0.4716645
mracemix	0.1932267
mracewhite	0.4245238
mpregwt_c	0.9895679
smoke1	1.3434559
mht_c	0.9720204

Table 4: Table of the 95% confident intervals for the coefficients beta_0 to beta_13 in the odds scale

	2.5 %	97.5 %
(Intercept)	0.1244872	7.0360889
med1	0.0933358	3.9460923
med2	0.0670341	2.7439018
med3	0.0729256	3.6722038
med4	0.0343270	1.5100938
med5	0.0549903	2.4332320
med7	0.3474163	118.1492597
mraceblack	0.3704402	2.1722449
mracemexican	0.1354003	1.6430354
mracemix	0.0215894	1.7293917
mracewhite	0.1887364	0.9548791
mpregwt_c	0.9791727	1.0000735
smoke1	0.9356378	1.9290304
mht_c	0.8954026	1.0551941

The intercept value indicates that a non smoker mother, with education category 0, of Asian race with mean values for the numeric variables, would have odds of 0.9358967 (log odds of -0.0662501) of having a premature birth vs. having a full-term birth on average. Mothers who smoke have odds of 1.3434559 of having a premature compared to having a full-term birth, although this is not statistically significant as seen by the odds interval of (0.9356378, 1.9290304) with 95% confidence. It includes odds = 1 in the interval, so it is likely that smoking does not have any effect in the chances of having a premature birth. It is likely with 95% confidence that the odds could change by any value between -6.43622% or +92.90304%.

For the full model in the stepwise selection process the interaction between smoke and race were taken into account. However, this interaction was not statistically significant and the null hypotheses that their coefficients were equal to zero could not be rejected. The following R code shows all the interactions between race and smoke are not significant and should not be included in the model as they do not affect the chances of having a premature birth: "test_model <- glm(formula = premature ~ med + mrace + mpregwt_c + smoke + mht + mrace:smoke, family = binomial(link = logit), data = smoking1) summary(test_model)".



There are no outliers on the binned residual plots for the numerial variables, which raises no additional concerns for the final model. Other than the variable of mracewhite no other variables or their interactions showed strong associations to the log odds of the probability of having a premature birth. Maybe other biological or genetic factors have a stronger impact on the birth weight, or maybe a new model taking into account the father's information would provide better results.

- **Annex**
  Annex 1.0 Final Model:

```
summary(model1)
```

```
##
## Call:
## glm(formula = premature ~ med + mrace + mpregwt_c + smoke + mht_c,
##     family = binomial(link = logit), data = smoking1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7366  -0.6715  -0.5585  -0.4070   2.4278
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.066250   1.029255  -0.064   0.9487
## med1         -0.499413   0.955190  -0.523   0.6011
## med2         -0.846587   0.946940  -0.894   0.3713
## med3         -0.658762   0.999791  -0.659   0.5100
## med4         -1.479825   0.965322  -1.533   0.1253
## med5         -1.005689   0.966808  -1.040   0.2982
## med7          1.857359   1.487063   1.249   0.2117
## mraceblack   -0.108651   0.451239  -0.241   0.8097
## mracemexican -0.751487   0.636763  -1.180   0.2379
## mracemix     -1.643891   1.118215  -1.470   0.1415
## mracewhite   -0.856787   0.413588  -2.072   0.0383 *
## mpregwt_c    -0.010487   0.005388  -1.946   0.0516 .
## smoke1        0.295245   0.184581   1.600   0.1097
## mht_c        -0.028379   0.041890  -0.677   0.4981
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 841.83  on 868  degrees of freedom
## Residual deviance: 795.45  on 855  degrees of freedom
## AIC: 823.45
##
## Number of Fisher Scoring iterations: 5
```

```
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(model1) >= 0.5, "1","0")),
                            as.factor(smoking1$premature),positive = "1")
Conf_mat$table
```

```
##           Reference
## Prediction   0   1
##          0 704 160
##          1   1   4
```

```
Conf_mat$overall["Accuracy"];
```

```
##  Accuracy
## 0.8147296
```

```
Conf_mat$byClass[c("Sensitivity","Specificity")] #True positive rate and True negative rate
```

```
## Sensitivity Specificity
##  0.02439024  0.99858156
```

```
#let's repeat with the marginal percentage in the data
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(model1) >= mean(smoking1$premature), "1","0")),
                            as.factor(smoking1$premature),positive = "1")
Conf_mat$table
```

```
##           Reference
## Prediction   0   1
##          0 442  66
##          1 263  98
```

```
Conf_mat$overall["Accuracy"];
```

```
##  Accuracy
## 0.6214039
```

```
Conf_mat$byClass[c("Sensitivity","Specificity")]
```

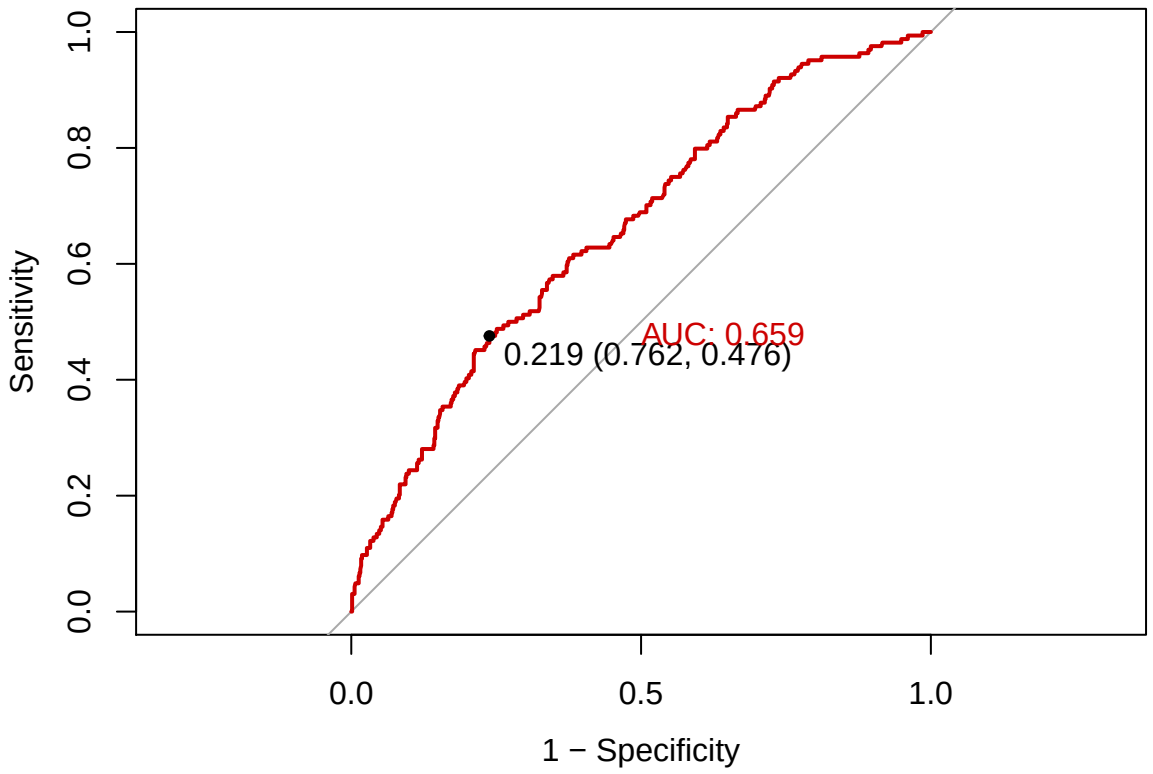
```
## Sensitivity Specificity
##   0.5975610   0.6269504
```

```
#still not moving much.... the model can predict only so well

#ROC curve...
roc(smoking1$premature,fitted(model1),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
```

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

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



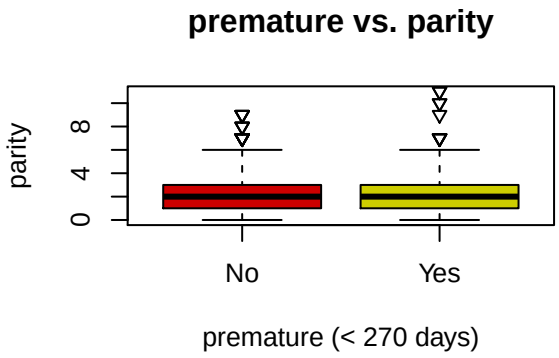
```
##
## Call:
## roc.default(response = smoking1$premature, predictor = fitted(model1),     plot = T, print.thres = "
##
## Data: fitted(model1) in 705 controls (smoking1$premature 0) < 164 cases (smoking1$premature 1).
## Area under the curve: 0.6593
```

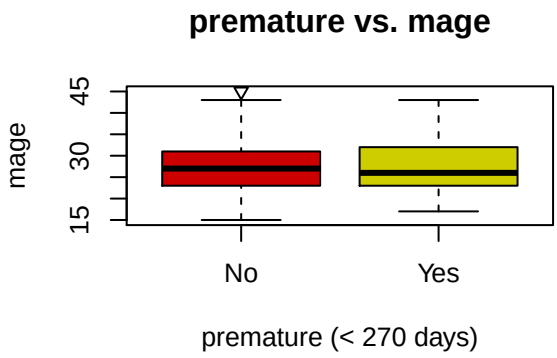
Annex 1.1 Data Summary:

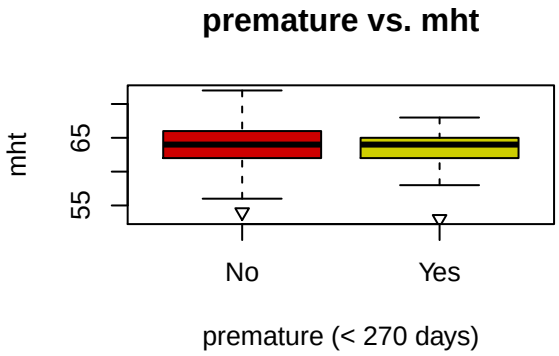
```
##      parity           mrace          mage       med          mht
##  Min.   : 0.000   asian  : 34   Min.   :15.00   0:  5   Min.   :53.00
##  1st Qu.: 1.000   black  :169   1st Qu.:23.00   1:130   1st Qu.:62.00
##  Median : 2.000   mexican: 25   Median :26.00   2:321   Median :64.00
##  Mean   : 1.953   mix    : 15   Mean   :27.29   3: 47   Mean   :64.07
##  3rd Qu.: 3.000   white  :626   3rd Qu.:31.00   4:203   3rd Qu.:66.00
##  Max.   :11.000                 Max.   :45.00   5:159   Max.   :72.00
##                                                 7:  4
##     mpregwt           inc      smoke   premature
##  Min.   : 87.0   1      :153   0:466   0:705
##  1st Qu.:113.0   2      :146   1:403   1:164
##  Median :125.0   3      :136
##  Mean   :128.5   7      :111
##  3rd Qu.:140.0   4      :105
##  Max.   :220.0   5      : 98
##                  (Other):120
```

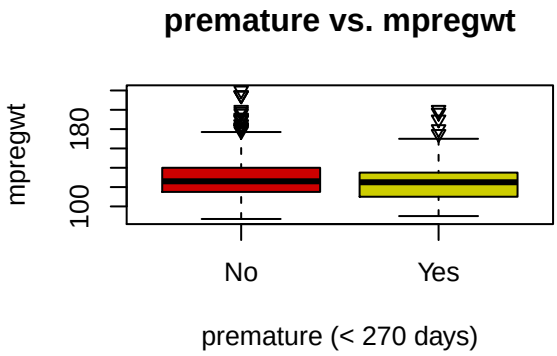
```
##             premature      parity         mage          mht     mpregwt
## premature  1.00000000  0.04494262  0.019428566 -0.093290040 -0.07654619
## parity     0.04494262  1.00000000  0.523690421 -0.042815618  0.15053789
## mage       0.01942857  0.52369042  1.000000000 -0.005470885  0.14613682
## mht       -0.09329004 -0.04281562 -0.005470885  1.000000000  0.46044630
## mpregwt   -0.07654619  0.15053789  0.146136818  0.460446304  1.00000000
```

Annex 1.2 Explotarory Data Analysis:







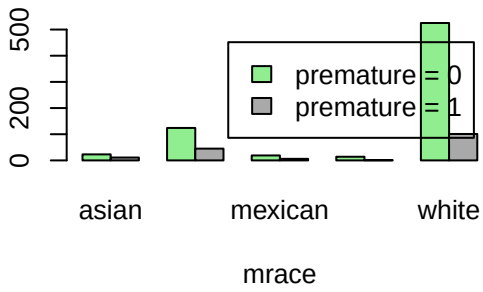


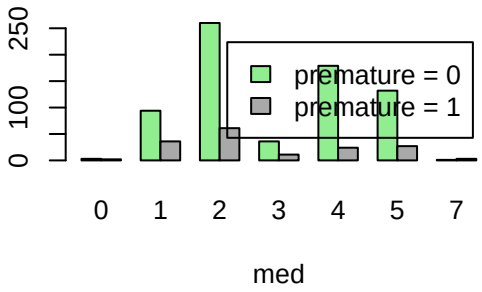

```
##
##         asian     black mexican        mix     white
##   0 0.6764706 0.7337278    0.76 0.93333333 0.8386581
##   1 0.3235294 0.2662722    0.24 0.06666667 0.1613419
```

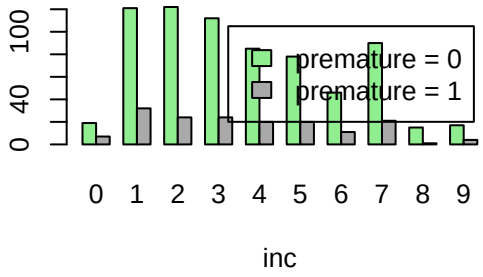
```
##
##       0         1         2         3         4         5    7
##   0 0.6 0.7230769 0.8099688 0.7659574 0.8817734 0.8301887 0.25
##   1 0.4 0.2769231 0.1900312 0.2340426 0.1182266 0.1698113 0.75
```

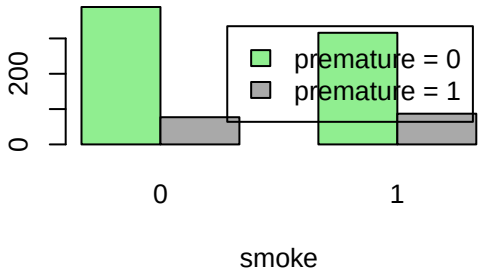
```
##
##             0         1         2         3         4         5         6
##   0 0.7307692 0.7908497 0.8356164 0.8235294 0.8095238 0.7959184 0.8070175
##   1 0.2692308 0.2091503 0.1643836 0.1764706 0.1904762 0.2040816 0.1929825
##
##             7      8         9
##   0 0.8108108 0.9375 0.8095238
##   1 0.1891892 0.0625 0.1904762
```

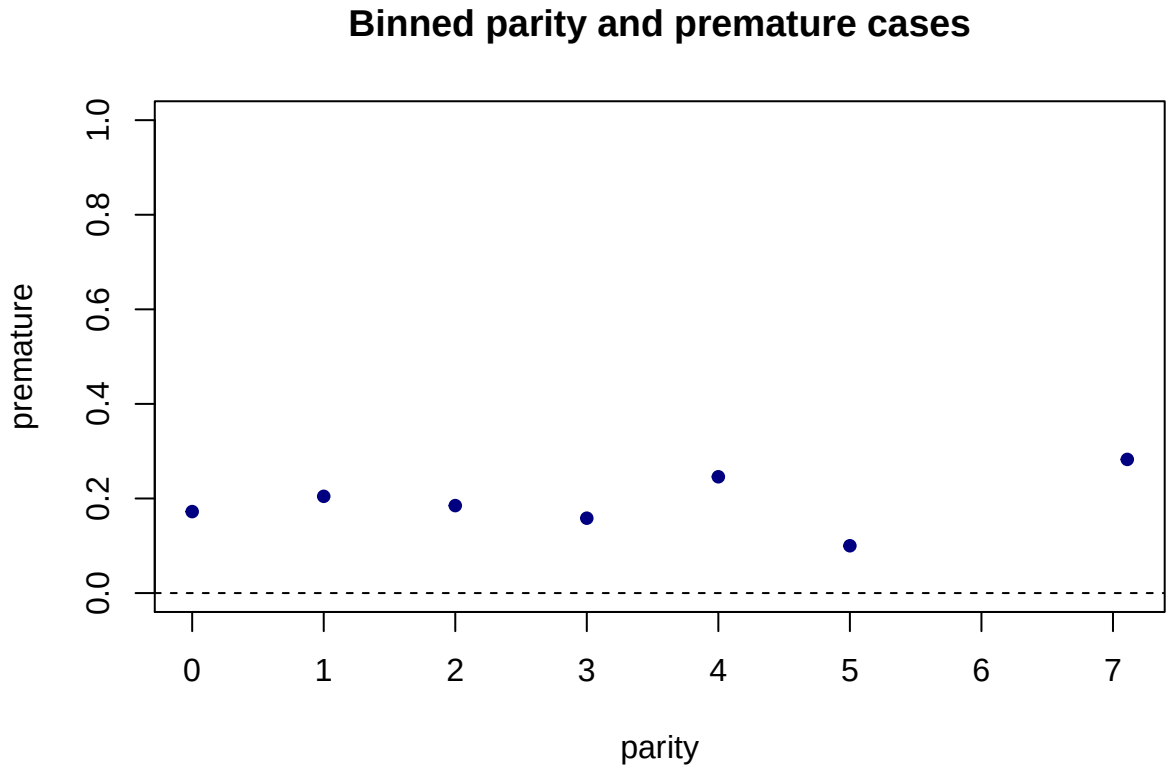
```
##
##             0         1
##   0 0.8347639 0.7841191
##   1 0.1652361 0.2158809
```

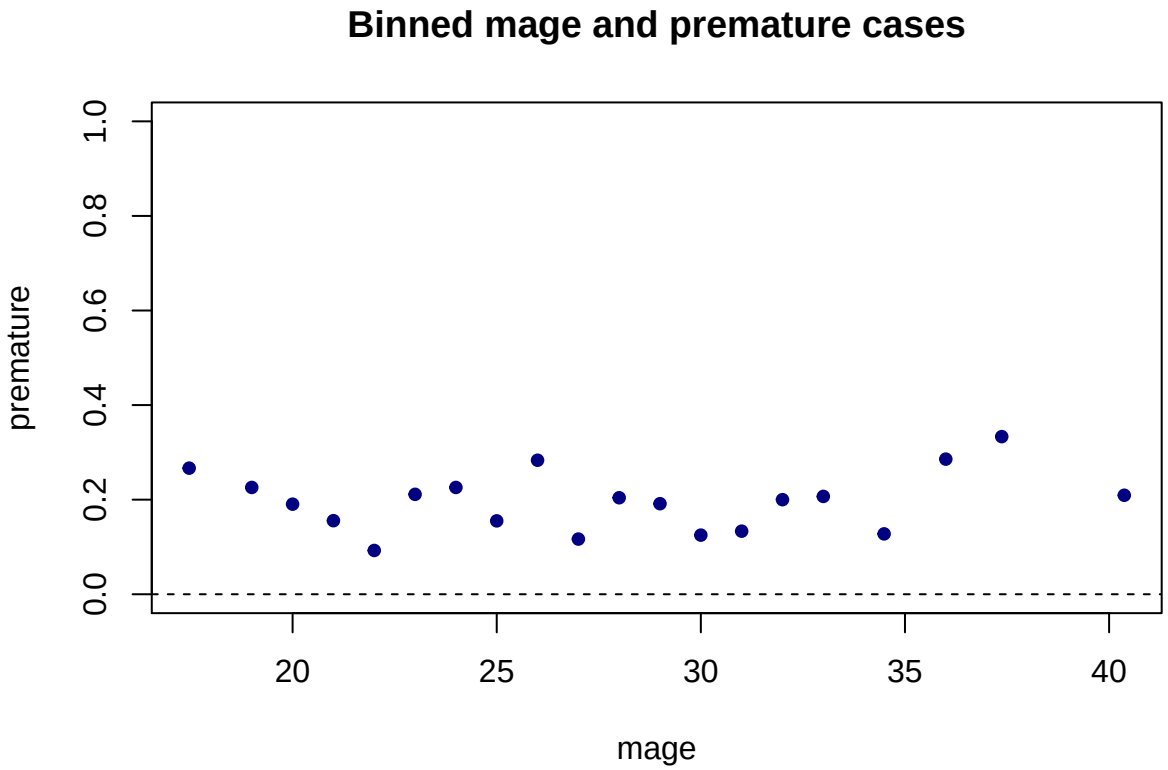

premature = 0
premature = 1
0
200
500
asian
mexican
white
mrace

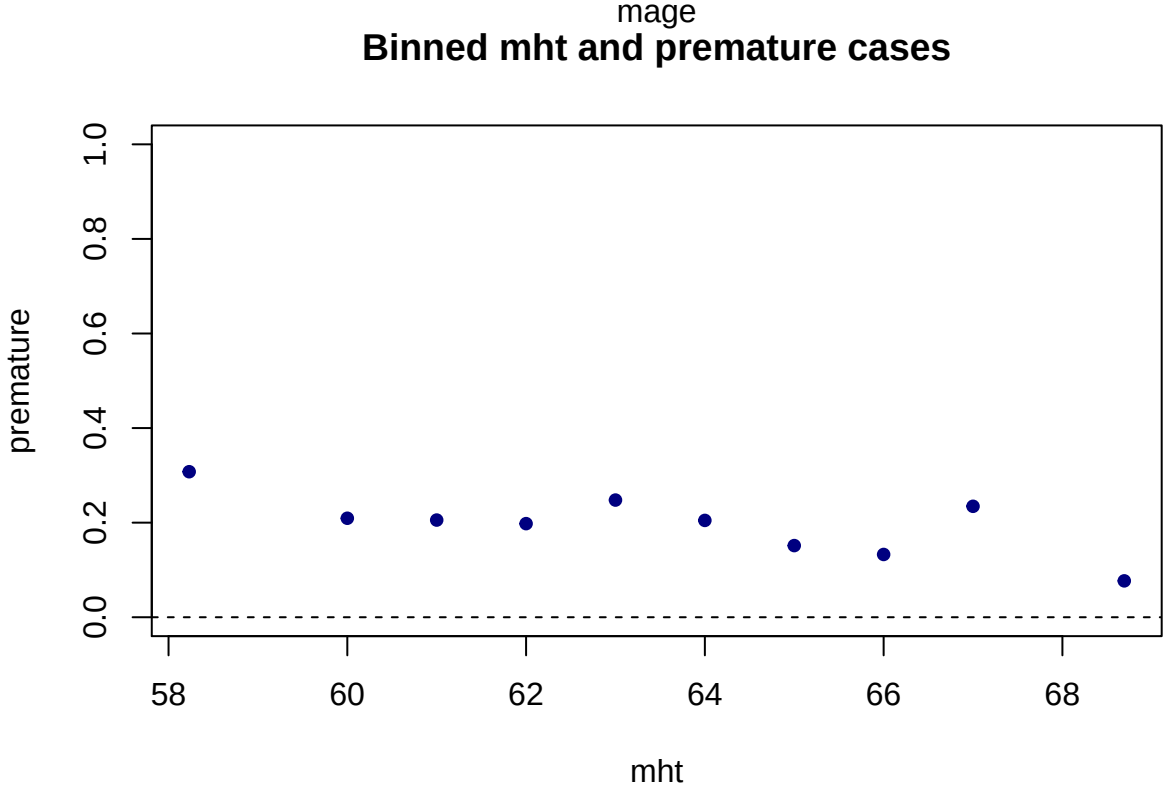

premature = 0
premature = 1
0
100
250
0
1
2
3
4
5
7
med

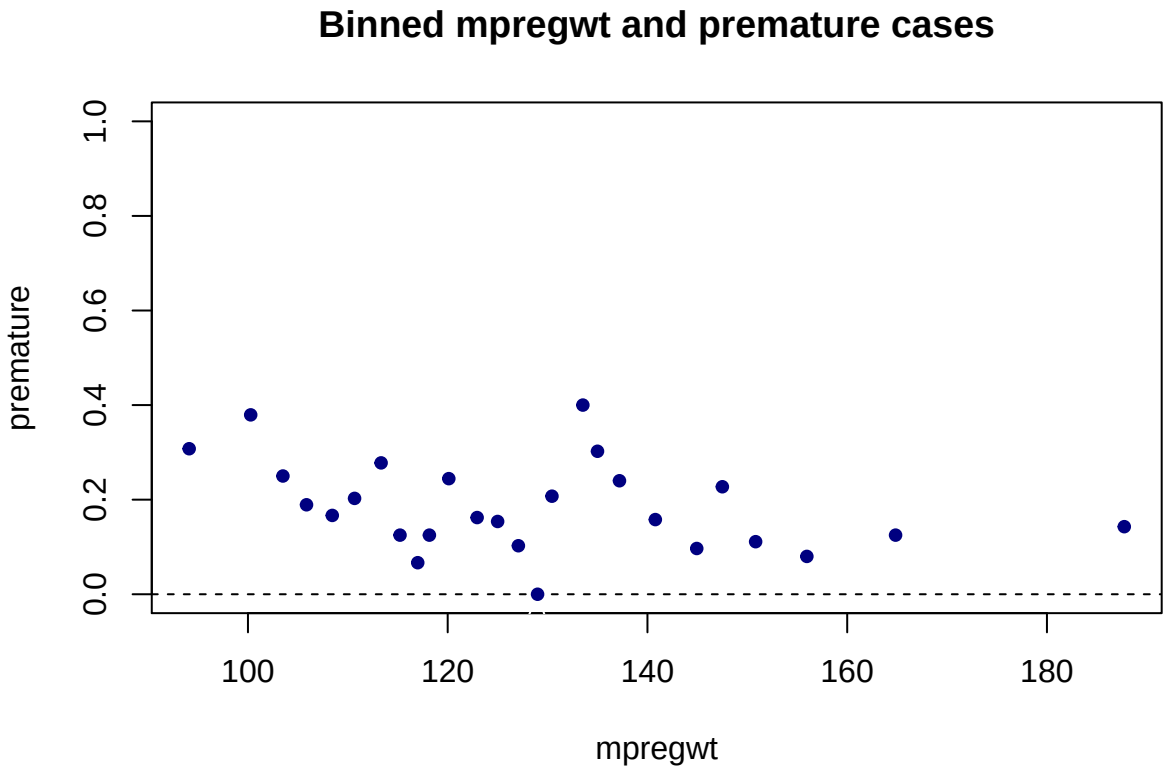

premature = 0
premature = 1
0
40
100
0
1
2
3
4
5
6
7
8
9
inc


premature = 0
premature = 1
0
200
0
1
smoke


Binned parity and premature cases
premature
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
parity


Binned mage and premature cases
premature
0.0
0.2
0.4
0.6
0.8
1.0
20
25
30
35
40
mage


Binned mht and premature cases
premature
0.0
0.2
0.4
0.6
0.8
1.0
58
60
62
64
66
68
mht


Binned mpregwt and premature cases
premature
0.0
0.2
0.4
0.6
0.8
1.0
100
120
140
160
180
mpregwt