Homework 3: Maternal Smoking and Pre-Term Birth

Guillem_Amat

September 26, 2019

Summary

After analyzing which factors cause Premature births in babies, it was concluded that mother's smoking habits and their race had a significant impact on premature births.

The odds ratio between mothers who smoked and those that did not, did not differ between races.

Mother's weight was another variable that had a significant impact on premature births.

Introduction

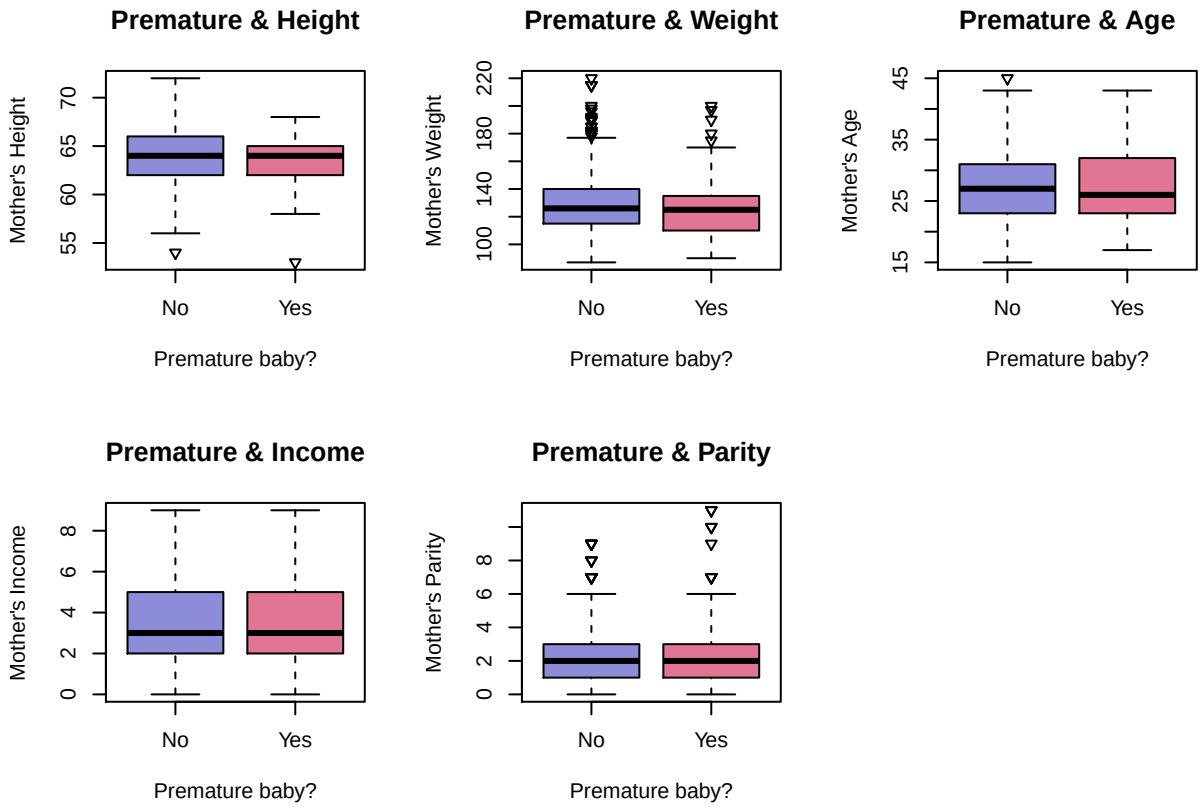
The study concerned investigating premature births in babies. The analysis targeted answering whether mothers who smoked tended to give birth to premature babies, whether the odds ratio between smokers and non-smokers differed between different races and whether there were any other interesting associations.

Data

The original dataset contained 21 variables that included parent's and pregnancy information for babies being born at the Kaiser Foundation Hospital. Many values where missing for the fathers, as it is sometimes hard to record information on them, so a first step in the analysis was to filter columns with this issue in order to get a dataset with complete observations.

Furthermore the 0-5 values in race were collapsed to white. The education levels were reorganized into more easily interpretable groups: Elementary School, High School, College and Trade School. The Id and date variables were excluded from the analysis as they did not add any specific information.

The continuous variables Height, Weight, Age, Parity and Income were plotted against Premature Births to check whether there could be an interesting association. While none of the variables seemed to have a large impact on Premature births, Parity and Income seemed to have no impact at all.



The categorical variables were also analyzed to see if there was an interesting association between them and Premature Births. Mother's who smoked seemed more likely to give birth to Premature babies (53% vs 47% in conditional probabilities). See table below:

```
##       Premature
## smoke         0         1
##     0 0.551773 0.4695122
##     1 0.448227 0.5304878
```

The number of Premature births per mother races was calculated. Certain mother races had larger numbers of Premature births, in particular babies born from black and asian mothers. Finally, the same was done for educational levels, and mothers that graduated from college had the lowest numbers of premature births.

Model

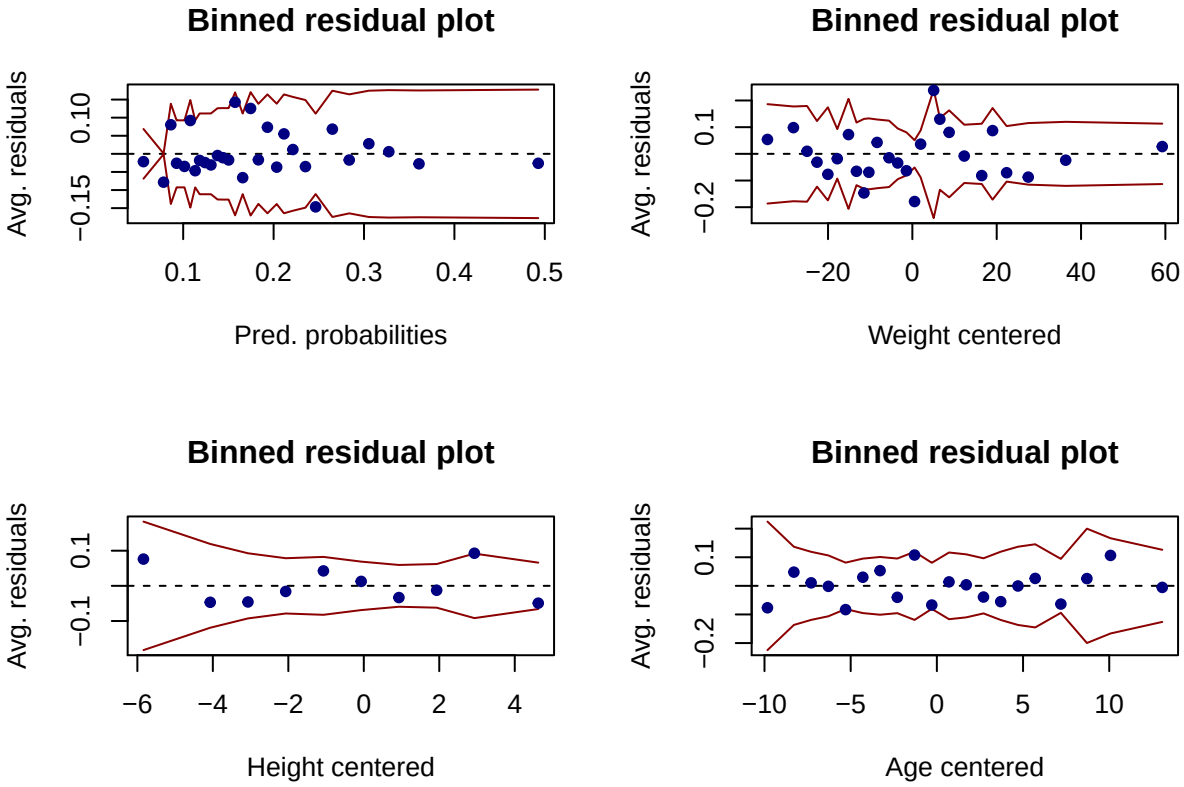
The final model selected was *Premature ~ med:mage_c + mht_c + mpregwt_c + med + smoke + mrace + mage_c*.

The model was selected based on what was learnt during the Exploratory Data Analysis:

- An initial logistic model with all the variables was fitted to the data.
- After that a second model without Income was fitted and a Chi-Squared test was performed between this second model and the first one to check whether Income had an influence on Premature births. Given the low p-value, we were not able to reject the null hypothesis and the variable was removed from the model. The same process was repeated for Parity, with the exact same results on the Chi-Squared test, so Parity was also removed.
- After that, Interactions were tested. A first interaction between race and smoke was tested by creating a model with it and performing a Chi-Squared test against the model without Income or Parity. Given

the low p-value, the interaction between mother's race and whether they smoked was dimmed not significant and was not included. The interaction between mother's education level and age was tested, which resulted in a low p-value, indicating a significant interaction, so it was included in the final model.

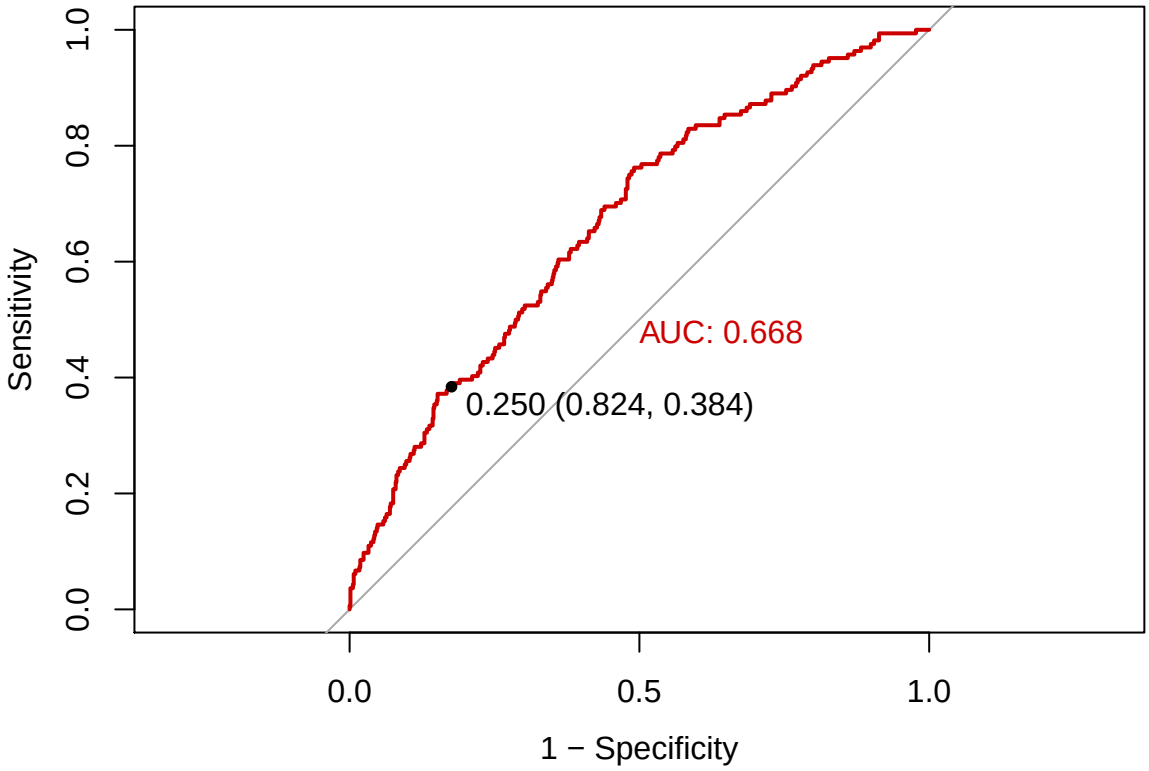
- Binned residuals plots were used to check that residuals were distributed randomly across positive and negative values. This was the case. This means that the model does a good job describing the data.



- Finally a Confusion Matrix and an ROC curve were used to diagnose how well the model performed. The AUC of the model was 0.668 and with a threshold of 0.25, the model's Accuracy was 74%, its Sensitivity 38% and its Specificity 82%.

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

## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = births$Premature, predictor = fitted(birthsmod4),     plot = T, print.thres =
##
## Data: fitted(birthsmod4) in 705 controls (births$Premature 0) < 164 cases (births$Premature 1).
## Area under the curve: 0.6683
```

Results and Conclusions

	x
(Intercept)	0.2351379
mht_c	0.9721707
mpregwt_c	0.9890683
medHigh School	0.7216158
medCollege	0.4769325
medTrade School	9.1360105
smoke	1.3737551
mracemexican	1.2214463
mraceblack	2.0388357
mraceasian	2.5709181
mracemix	0.4112797
mage_c	0.9722322
medHigh School:mage_c	1.0868686
medCollege:mage_c	1.0169008
medTrade School:mage_c	1.1222993

	2.5 %	97.5 %
(Intercept)	0.1415461	0.3906136
mht_c	0.8952614	1.0556872
mpregwt_c	0.9784883	0.9997627
medHigh School	0.4427408	1.1761493
medCollege	0.2825593	0.8050154
medTrade School	0.8329022	100.2118675
smoke	0.9546911	1.9767682
mracemexican	0.4389403	3.3989386
mraceblack	1.3193808	3.1506073
mraceasian	1.1399823	5.7980021
mracemix	0.0521856	3.2413363
mage_c	0.9167112	1.0311160
medHigh School:mage_c	1.0107879	1.1686758
medCollege:mage_c	0.9309906	1.1107386
medTrade School:mage_c	0.6815467	1.8480843

The Intercept represents the odds of giving birth to a premature baby for a white mother that does not smoke, that only has elementary or middle school education and that is of average height, age and weight. The odds of this situation were 0.23, and we are 95% confident that the value of the Intercept fell between 0.14 and 0.39.

Mothers with the aforementioned characteristics but that also smoked were 1.37 times more likely to give birth to premature babies. The 95% confidence interval in this case was 0.95 and 1.97.

In terms of race, black mothers were 2.03 times more likely to give birth to premature babies (with 95% confidence interval between 1.3 and 3) and asian mothers 2.57 more likely (with confidence interval between 1.13 and 5.7). Race seemed to be a good predictor of premature births.

Finally, I also wanted to note the predictive power of mother's weight, which had a low p-value, indicating a relationship with premature births. No other variables in the dataset other than the ones mentioned seemed to be great predictors of premature births.

Analysis and Model Limitations

There were some potential limitations in the analysis, specifically coming from the dataset. There was a way larger number of normal births than premature ones. Furthermore, the races were not distributed evenly, ~70% of the mothers were white. A more complete dataset could help improve the predictive capacity of the model.

It would also be interesting to include other variables in the analysis, such as a variable that captured mother's habits during the pregnancy.

Appendix

Model Creation

Centering the continuous variables to aid in interpretation

```
births$mpregwt_c <- births$mpregwt - mean(births$mpregwt)
births$mht_c <- births$mht - mean(births$mht)
births$mage_c <- births$mage - mean(births$mage)

birthsmod1 <- glm(Premature ~ mht_c + mpregwt_c + med + smoke + mrace + mage_c + parity + inc, data = b
summary(birthsmod1)
```

```
##
## Call:
## glm(formula = Premature ~ mht_c + mpregwt_c + med + smoke + mrace +
##     mage_c + parity + inc, family = binomial, data = births)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7522  -0.6811  -0.5485  -0.4290   2.3684
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.363018   0.332348  -4.101 4.11e-05 ***
## mht_c           -0.029427   0.041953  -0.701  0.48303
## mpregwt_c       -0.010890   0.005476  -1.989  0.04673 *
## medHigh School  -0.423678   0.254127  -1.667  0.09548 .
## medCollege      -0.891214   0.282099  -3.159  0.00158 **
## medTrade School  2.261860   1.184881   1.909  0.05627 .
## smoke            0.297009   0.184508   1.610  0.10746
## mracemexican     0.156025   0.521234   0.299  0.76468
## mraceblack       0.741392   0.228533   3.244  0.00118 **
## mraceasian       0.896142   0.411804   2.176  0.02955 *
## mracemix        -0.842267   1.050692  -0.802  0.42277
## mage_c           0.019177   0.019833   0.967  0.33357
## parity          -0.026728   0.059344  -0.450  0.65243
## inc              0.024651   0.042802   0.576  0.56466
## ---
## 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: 796.48  on 855  degrees of freedom
## AIC: 824.48
##
## Number of Fisher Scoring iterations: 5
```

Testing removing Parity and Income

```
birthsmod2 <- glm(Premature ~ mht_c + mpregwt_c + med + smoke + mrace + mage_c + inc, data = births, fa
anova(birthsmod1, birthsmod2, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ mht_c + mpregwt_c + med + smoke + mrace + mage_c +
##     parity + inc
## Model 2: Premature ~ mht_c + mpregwt_c + med + smoke + mrace + mage_c +
##     inc
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       855     796.48
## 2       856     796.69 -1 -0.20383   0.6516
```

```
birthsmod3 <- glm(Premature ~ mht_c + mpregwt_c + med + smoke + mrace + mage_c + parity, data = births,
anova(birthsmod1, birthsmod3, test= "Chisq")
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: Premature ~ mht_c + mpregwt_c + med + smoke + mrace + mage_c +
##     parity + inc
## Model 2: Premature ~ mht_c + mpregwt_c + med + smoke + mrace + mage_c +
##     parity
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       855     796.48
## 2       856     796.82 -1 -0.33096   0.5651
```

Checking whether there is a relation between mother's smoking and race

```
birthsmod4 <- glm(Premature ~ mrace:smoke + mht_c + mpregwt_c + med + smoke + mrace + mage_c, data = bi
anova(birthsmod1, birthsmod4, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ mht_c + mpregwt_c + med + smoke + mrace + mage_c +
##     parity + inc
## Model 2: Premature ~ mrace:smoke + mht_c + mpregwt_c + med + smoke + mrace +
##     mage_c
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       855     796.48
## 2       853     791.95  2   4.5356   0.1035
```

Checking whether there is a relation between mother's Age and Mother's eduaction level

```
birthsmod4 <- glm(Premature ~ med:mage_c + mht_c + mpregwt_c + med + smoke + mrace + mage_c, data = birt
anova(birthsmod1, birthsmod4, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ mht_c + mpregwt_c + med + smoke + mrace + mage_c +
##     parity + inc
## Model 2: Premature ~ med:mage_c + mht_c + mpregwt_c + med + smoke + mrace +
##     mage_c
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       855     796.48
## 2       854     790.83  1   5.6545  0.01741 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary of the Model

```
summary(birthsmod4)
```

```
##
## Call:
## glm(formula = Premature ~ med:mage_c + mht_c + mpregwt_c + med +
##     smoke + mrace + mage_c, family = binomial, data = births)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7332  -0.6793  -0.5359  -0.4112   2.3634
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.447583   0.258957  -5.590 2.27e-08 ***
## mht_c                  -0.028224   0.042050  -0.671  0.50209
## mpregwt_c              -0.010992   0.005487  -2.003  0.04515 *
```

```
## medHigh School         -0.326262   0.249243  -1.309  0.19053
## medCollege             -0.740380   0.267090  -2.772  0.00557 **
## medTrade School         2.212224   1.221993   1.810  0.07024 .
## smoke                   0.317548   0.185675   1.710  0.08722 .
## mracemexican            0.200036   0.522167   0.383  0.70165
## mraceblack              0.712379   0.222053   3.208  0.00134 **
## mraceasian              0.944263   0.414931   2.276  0.02286 *
## mracemix               -0.888482   1.053319  -0.844  0.39895
## mage_c                 -0.028161   0.030002  -0.939  0.34792
## medHigh School:mage_c   0.083301   0.037026   2.250  0.02446 *
## medCollege:mage_c       0.016760   0.045034   0.372  0.70978
## medTrade School:mage_c  0.115380   0.254479   0.453  0.65026
## ---
## 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: 790.83  on 854  degrees of freedom
## AIC: 820.83
##
## Number of Fisher Scoring iterations: 5
```