Methods and Data Analysis 3

Srishti Saha (ss1078)

27 September, 2019

Summary

This problem is based on the maternal smoking and birth weigths data. However, the objective is to investigate and establish the relationship between gestational age and other variables. The variable gestation is also recoded as 'Premature' which is basically a binary variable the follows the logic: gestational age<270 is counted as Premature (equal to 1) and 0 otherwise. This problem is thus a problem of logistic regression.

Introduction

This report covers the different data preprocessing and transformation steps, the process of exploratory data analysis and the exercise of determining the relationship of different variables with the babies' gestational age. The objective of the analysis is based on the Surgeon Generals' claim that mothers who smoke have increased rates of premature delivery (before 270 days) and low birth weights. The analysis will answer the following questions:

- 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

Data Dictionary

We will now define the variable gestation and Premature and do a quick check on the distribution of data across Premature =1 and =0. The definition of the variables are: * **gestation**- *length of Gestation period of the child*; * **Premature**- *If gestational age < 270 days- Premature=1, else 0*

The distribution of data across Premature levels (0 and 1) is: 164 premature births and 705 normal births. The distribution of data across smoke levels (0 and 1) is: 403 smoking moms and 466 non-smoking moms.

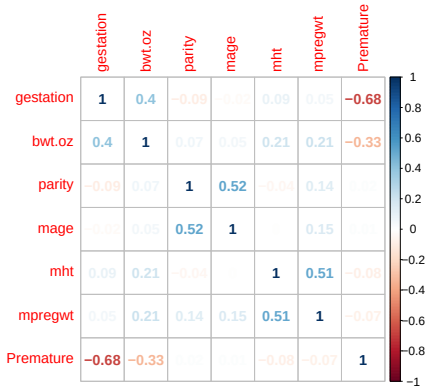
Data Transformation

We have already performed a few data transformations, namely: >1. converted education, race, income and smoke to factor variables 2. In race: collapsed levels 0-5 to '5' indicating 'White' as per instructions 3. In education: collapsed levels 6-7 to '7' indicating 'trade school' as per instructions (however, there were no records for med=6) 4. We also dropped the ID and date columns owing to the nature of the variables. These variables cannot be used for predictions or establishing associations.

```
## [1] "Resultant dimensions of the dataset are: 869 rows and 11 columns"
```

EDA

Let us first plot the correlation of these variables.



From the correlation plot, we see that there is a relatively high correlation (0.51) between the mothers' height and pregnancy weight. As explained in the question, birth weight of the baby and gestation. Since gestation is an indirect outcome variable (as it determines prematurity), we will be not be selecting birth weight in the analysis. Since Premature is a derived variable from gestation, we will also be removing that eventually. We will now move on to the categorical variables.

We see that there is no strong relationship between the smoking indicator and the premature flag. A lot of mothers (78%) who smoked have had normal births i.e. 316/403. Moreover, 17% of mothers who do not smoke have had premature births (which is not very different from the 21% mark of mothers who smoke and have had a premature birth). Let us also look at the conditional probabilities of smoking and premature birth.

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

It seems like the conditional probability of premature birth given the mother smokes is 21.5%. However, the conditional probability of premature birth given the mother does not smoke is 16.5%. This does not give us any clear indication. It is interesting to note that the conditional probability of no premature birth given the mother does not smoke is 83.5%.

Let us also look at the relation between gestation and smoke.

Relation between smoke and premature: Here the null hypothesis is **Ho: there is no association between the two variables**. Since the p-value is high, we should consider the null hypothesis. Since the p-value is not very high as compared to the significance level of 0.05, it is on the border line and should be investigated through the model. **Relation between race and premature**: Since the p-value is low (0.003), we scan reject the null hypothesis. There seems to be a relation between race and premature. A similar result has been shown by the violin plot of gestation versus race in smoking and non-smoking mothers. **Relation between**

race and education: Since the p-value is low (0.0005), we scan reject the null hypothesis. There seems to be a relation between race and premature. A similar result has been shown by the violin plot of gestation versus race in smoking and non-smoking mothers. **Relation between race and income and race and parity**: While income seems to be insignificant due to the high p-value of the chi-squared test, parity comes out to be slightly associated with Premature due to its p-value of 0.01 (<0.05). We may need to investigate these variables.

Also checking binned plots, we have the following observations: > 1. The relationship between weight and premature cases does not have any pattern. The data is scarce towards the higher levels of weight. This might be because of lack of sufficient data. > 2. The relationship between age and premature cases does not have any pattern and is very linear. There is no evident pattern > 3. The relationship between height and premature cases does not have any pattern. It is a very linear pattern

> For all above relationships, probability does not start to decrease or shows a very slight decrease in weight. Since we mentioned it may be because of sparse data, so it is unlikely we would want a quadratic term. We would expect some flattening with a linear trend.

The other plots and result tests have been included in the appendix.

Model

The first model was with main effect for every variable and linear predictors. We also mean centred the continuous predictors.

With the first model with all base variables, we see that only race level 7 seee=ms to be significant with a low p-value (0.0007). We see that none of the levels of income are significant. The chi-squared test of Pramture with income also revealed that income is not a significant predictor. We can start with removing that. Let us also remove parity as only parity level 5 comes out to be very mildly significant. None of the other levels are significant based on their high p-values.For education, only level 7 seems to be mildly significant based on its p-value. However, since the chi-squared test revealed its association with Premature, we will retain it for now.

This model (model with base effects of mage_c, mht_c, mpregwt_c, med, mrace, smoke) shows that race (level 7 and level 8) are significant. mpregwt is significant too. Let us do a step-wise analysis to check the significant of the variables in this model once. The step-wise analysis using AIC showed that the mothers' weight is a significant variable along with race, education, and the smoke variable. Since the correlation matrix showed high correlation between height and weight, we remove height. Moreover, the variable for age was removed. We also saw from the EDA that age had a very low correlation with gestation and not a significant relationship with Premature either. Thus we will remove age. Creating a new model with main effects of: mpregwt_c, med, mrace, smoke

This model has an increased significance of mpregwt based on its z-score (from 2.054 to 2.508). The other two significant variables are mrace for level 7 and 8. Let us now try introducing interaction variables. We introduced interaction variables between the three significant main levels: race, education and mothers' weight. In this model, none of the interaction variables are significant. However, let us retain the interaction variable between race and smoke and create a new model.

Table 2: Model Comparison

Model.Name	Accuracy	Sensitivity	Specificity	AUC
model1	62.03%	62.20%	61.99%	66.7%
model2	61.68%	57.93%	62.55%	66.1%
model3	60.87%	57.93%	61.56%	65.7%
model4	60.18%	58.54%	60.57%	66.7%
model5	60.18%	60.98%	60.00%	66.4%

term	estimate	std.error	statistic	p.value
(Intercept)	-0.9236883	0.9568314	-0.9653616	0.3343637
mpregwt_c	-0.0126847	0.0048332	-2.6245005	0.0086776
med1	-0.5591958	0.9638881	-0.5801460	0.5618162
med2	-0.9064271	0.9601757	-0.9440221	0.3451584
med3	-0.7412153	1.0144015	-0.7306922	0.4649672
med4	-1.5744080	0.9738986	-1.6166036	0.1059639
med5	-1.0631724	0.9768641	-1.0883524	0.2764396
med7	1.8392054	1.5061572	1.2211245	0.2220389
mrace6	0.1873910	0.6292083	0.2978203	0.7658403
mrace7	1.0551947	0.3058067	3.4505285	0.0005595
mrace8	0.8273303	0.4946915	1.6724166	0.0944422
mrace9	-13.5149544	413.9504623	-0.0326487	0.9739547
smoke1	0.3971004	0.2278978	1.7424497	0.0814298
mrace6:smoke1	-0.0324753	1.1125336	-0.0291904	0.9767127
mrace7:smoke1	-0.5652227	0.4240751	-1.3328365	0.1825855
mrace8:smoke1	0.3170296	0.8451212	0.3751292	0.7075644
mrace9:smoke1	14.4624238	413.9524020	0.0349374	0.9721297

For the new model, let us do an anova test i.e. change in deviance test with the model with only the main effects of the individual variables. The results do not show any significant improvement based on the high p-value (0.27).

The sensitivity of the model with only the main effects of the variables: mpregwt_c, med, mrace and smoke is 57.9% while the one with the interaction term between mothers' race and smoke has a sensitivity of 60.9%. This is relevant because the model helps capture the cases for actual premature births more accurately. Let us also compare the ROC curves.

The AUC of the model without interaction term is 0.657 and the AUC of the model with the interaction term is 0.664. Based on these observations, my final model considers the following predictors: centred weight, education, race, smoke and interaction term between smoke and race.

Results

Final Model: Predictors are centred weight, education, race, smoke and interaction term between smoke and race

1. In the final model, the significant variables on the basis of their p-values are: centered weight, race (levels 7 and 8) and smoke (level 1). Of these, the most significant variable on the basis of the absolute value of z-score is race level 7 followed by mothers' weight.

2. The model estimate for mrace (level=7) is 1.06 on the log-scale which converts to 2.88 on the exponential scale. This means that the odds for a premature birth increases 2.88 times for a mother belonging to race level 7 (Black) as compared to level 5 (White).
3. The model estimate for mpregwt_c (centered weight) on a log-scale is -0.012 which indicates that for an increase in the mothers' weight by 1 pound, there will be a $e^{-0.012}$ (or 0.988) times decrease in the odds of premature birth.
4. None of the interaction terms with race and smoke are significant. This means that odds ratio of pre-term birth for smokers and non-smokers does not by mother's race.
5. The intercept estimate is -0.92 (although does not come out to be significant) on the log scale (i.e. -0.398) which means that **the odds ratio of prematurity given that all other variables are 0 and weight is centered at mean will be 40%**

As we see that the AUC is 66.4%. Moreover, the accuracy of the model is 60.18%. The sensitivity or the true-positive rate is 60.98%. The specificity or the true negative rate is 60%.

From the binned residual plot, we see that there is exactly one point on the confidence interval and one point outside it.

Conclusion

1. 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? **In the chosen model-model5 the coffecient is 0.397. Thus the log odds that if for smokers increases the probability of prematurity by 39.7%. The increase in odds ratio is exp(0.397) which is equal 1.49. This implies that the chance of having a premature baby increases by a factor of 1.49 or 49% if smoke = 1. The confidence interval for odds ratio of smoke (level 1) is [0.95,2.32]**
2. 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. **The used moel incorporates the interaction term between race and smoke. However, due to the high p-values, we can conclude that the interaction terms are not significant. Thus, the odds ratio of pre-term birth for smokers and non-smokers do not differ by mother's race**
3. Are there other interesting associations with the odds of pre-term birth that are worth mentioning? In the final model model5, other than smoke, we have used 4 terms: **med**, **mpregwt**, **mrace** and interaction term- **smoke:race**. However, the interaction term is not significant. The interpretation for all other significant variables are included in model results.

Potential Limitations

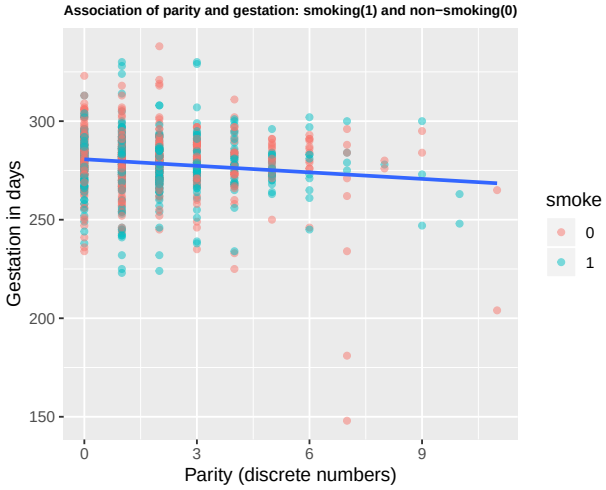
1. The data is imbalanced. It is skewed towards non-premature births with only 18% of the data represents Premature = 1.
2. The dtaa is also imbalanced when it comes to the race variable. It does not contain enough datapoints for all races apart from the baseline group i.e. American. This might lead to unreliable results of the interaction between Smoke and Race.
3. The chi-square tests were throwing warnings implying that the results might not be accurate due to lesser number of data points.

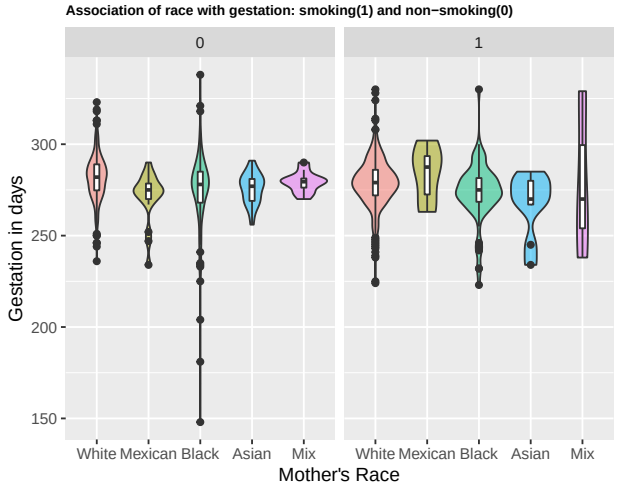
Appendix

Chi-squared test for smoke and premature

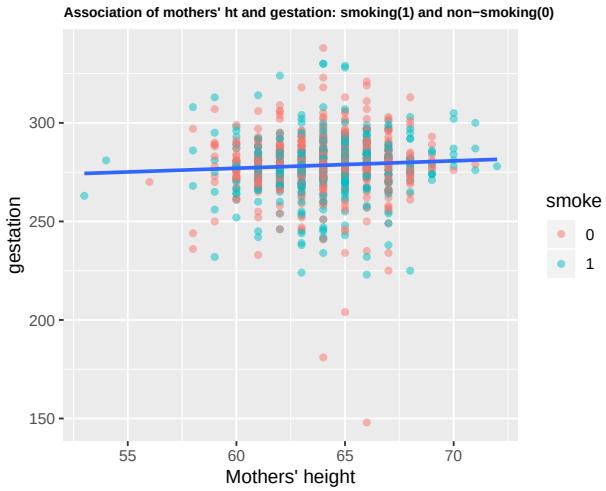
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(smoking_filtered[, c("smoke", "Premature")])
## X-squared = 3.2971, df = 1, p-value = 0.0694
```

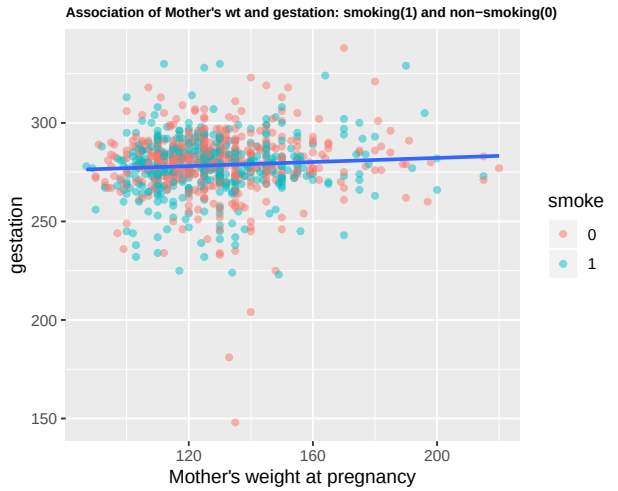
Plots to show association of categorical variables with gestation perios across smoking and non-smoking mothers





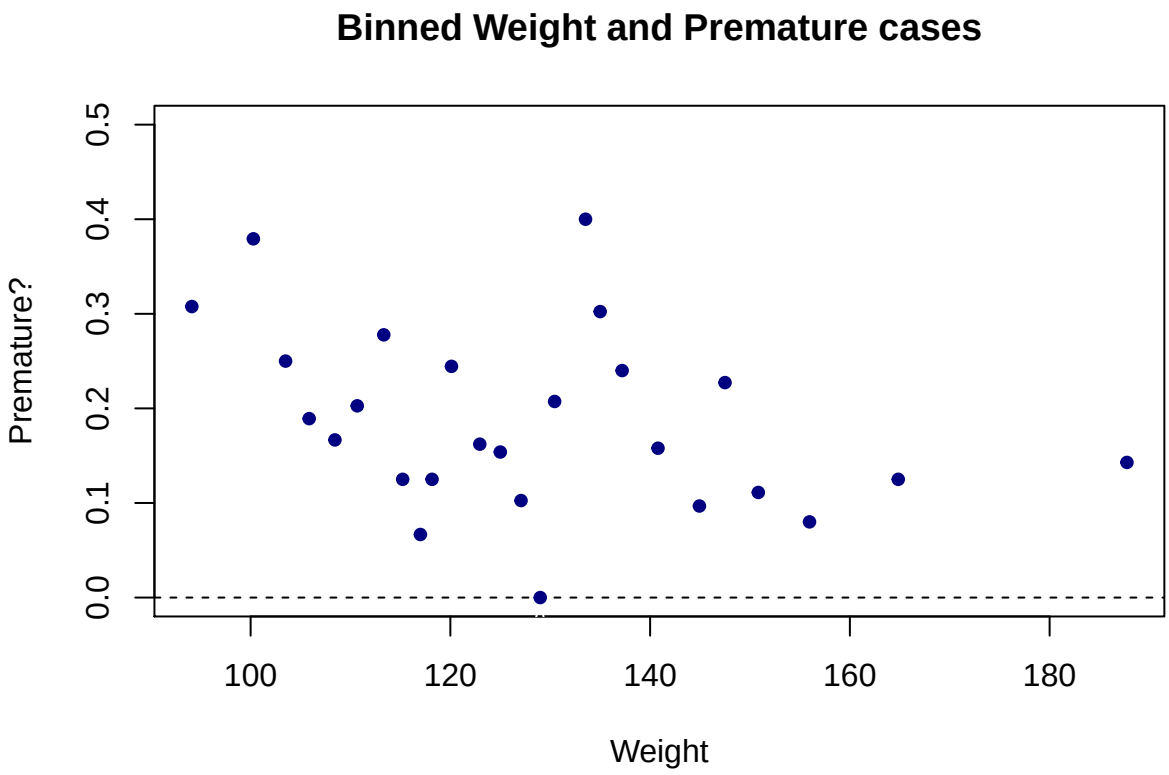
Plots to show association of numerical variables with gestation period across smoking and non-smoking mothers



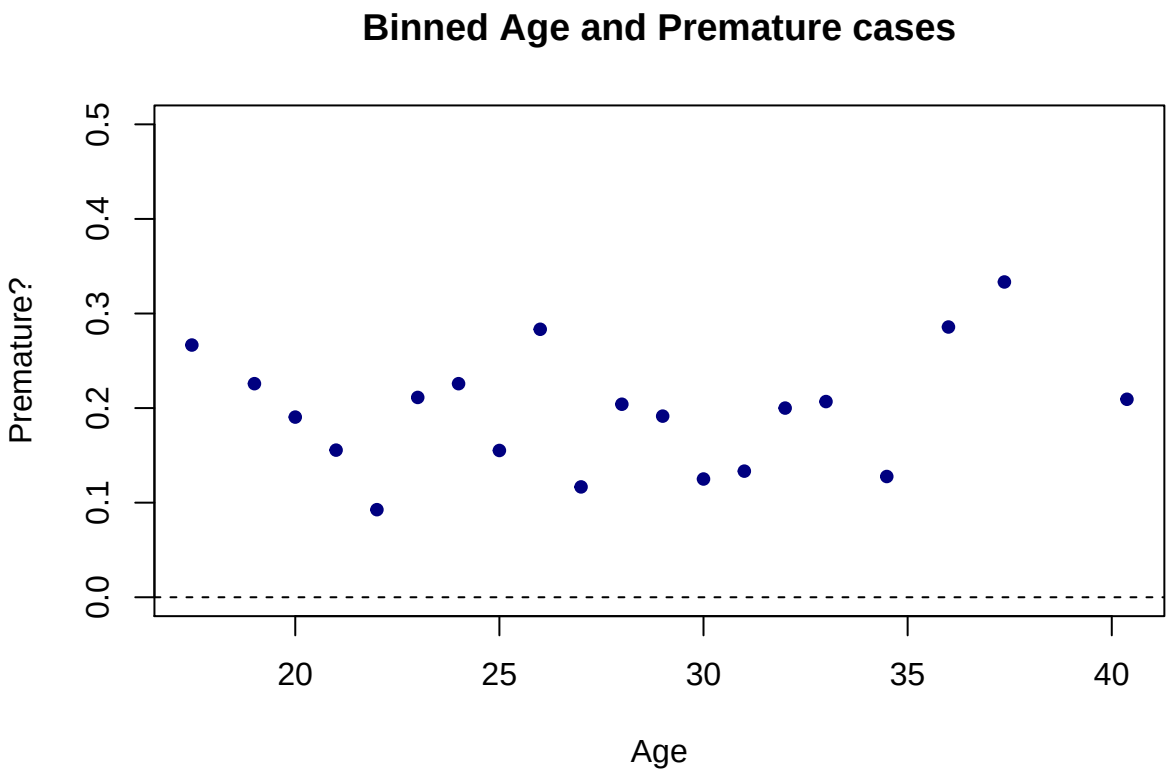


Binned Plots for EDA

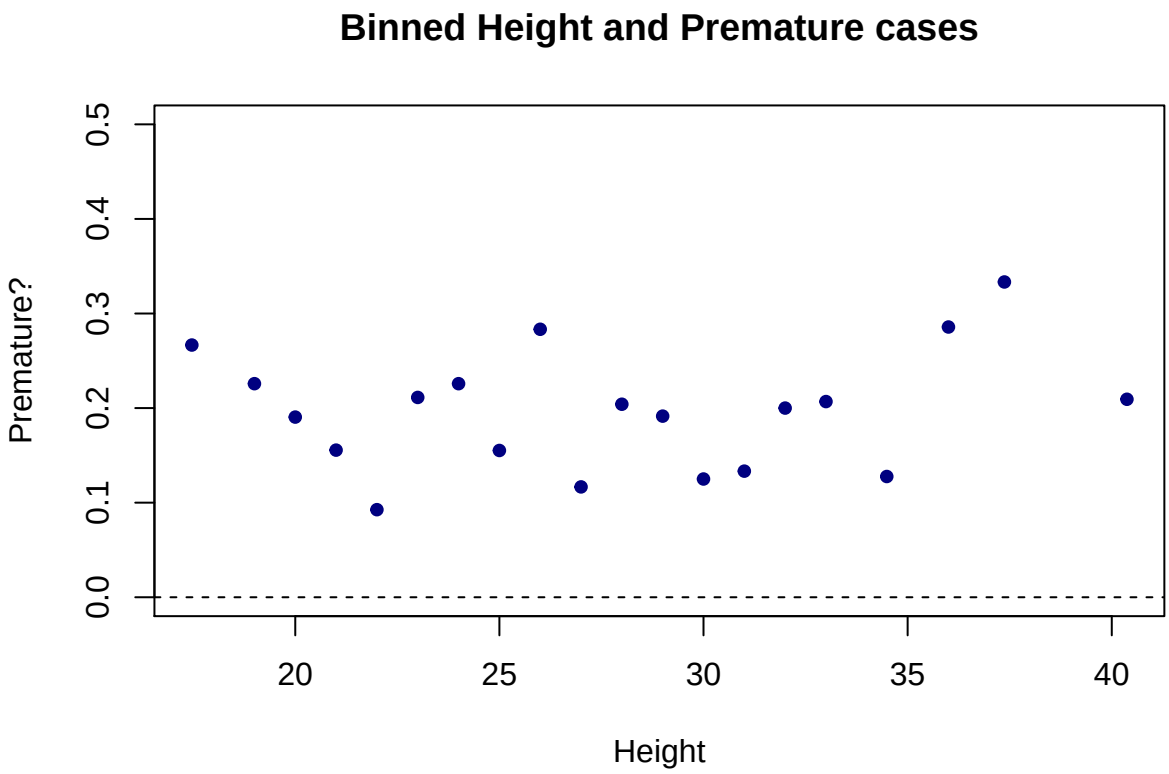
Weight



Age


Binned Age and Premature cases
Premature?
0.0
0.1
0.2
0.3
0.4
0.5
20
25
30
35
40
Age

Height



Chi-squared tests

With race

```
## Warning in chisq.test(table(smoking_filtered[, c("mrace", "Premature")])):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(smoking_filtered[, c("mrace", "Premature")])
## X-squared = 15.628, df = 4, p-value = 0.003561
```

With parity

```
## Warning in chisq.test(table(smoking_filtered[, c("parity", "Premature")])):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(smoking_filtered[, c("parity", "Premature")])
## X-squared = 24.372, df = 11, p-value = 0.01125
```

With income

```
## Warning in chisq.test(table(smoking_filtered[, c("inc", "Premature")])):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(smoking_filtered[, c("inc", "Premature")])
## X-squared = 4.0413, df = 9, p-value = 0.9087
```

With education

```
## Warning in chisq.test(table(smoking_filtered[, c("med", "Premature")])):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(smoking_filtered[, c("med", "Premature")])
## X-squared = 23.888, df = 6, p-value = 0.0005476
```

Model 1 (mage_c + mht_c + mpregwt_c + med + mrace + inc + parity + smoke)- summary

Model summary is:

```
##
## Call:
## glm(formula = Premature ~ mage_c + mht_c + mpregwt_c + med +
##     mrace + inc + parity + smoke, family = binomial, data = smoking_filtered)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7387  -0.6764  -0.5500  -0.3881   2.5343
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.674935   1.035890  -0.652 0.514691
## mage_c       0.012856   0.020611   0.624 0.532791
## mht_c       -0.033760   0.042919  -0.787 0.431522
## mpregwt_c   -0.010863   0.005554  -1.956 0.050504 .
## med1        -0.282620   0.986102  -0.287 0.774416
## med2        -0.675080   0.972915  -0.694 0.487762
## med3        -0.566646   1.019745  -0.556 0.578434
## med4        -1.335883   0.989002  -1.351 0.176780
## med5        -0.887571   0.990729  -0.896 0.370318
## med7         2.101549   1.503982   1.397 0.162316
## mrace6       0.111798   0.527851   0.212 0.832264
## mrace7       0.788333   0.235826   3.343 0.000829 ***
## mrace8       0.811621   0.418607   1.939 0.052518 .
## mrace9      -0.706575   1.058301  -0.668 0.504357
```

```
## inc1        -0.518715   0.514062  -1.009 0.312950
## inc2        -0.661162   0.523301  -1.263 0.206429
## inc3        -0.348798   0.526982  -0.662 0.508049
## inc4        -0.341002   0.537416  -0.635 0.525740
## inc5        -0.219518   0.541463  -0.405 0.685171
## inc6        -0.305657   0.586900  -0.521 0.602507
## inc7        -0.369093   0.539610  -0.684 0.493976
## inc8        -1.481360   1.152435  -1.285 0.198647
## inc9        -0.143388   0.753439  -0.190 0.849065
## parity      -0.005572   0.060543  -0.092 0.926671
## smoke1       0.302895   0.186150   1.627 0.103705
## ---
## 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: 790.21  on 844  degrees of freedom
## AIC: 840.21
##
## Number of Fisher Scoring iterations: 5
```

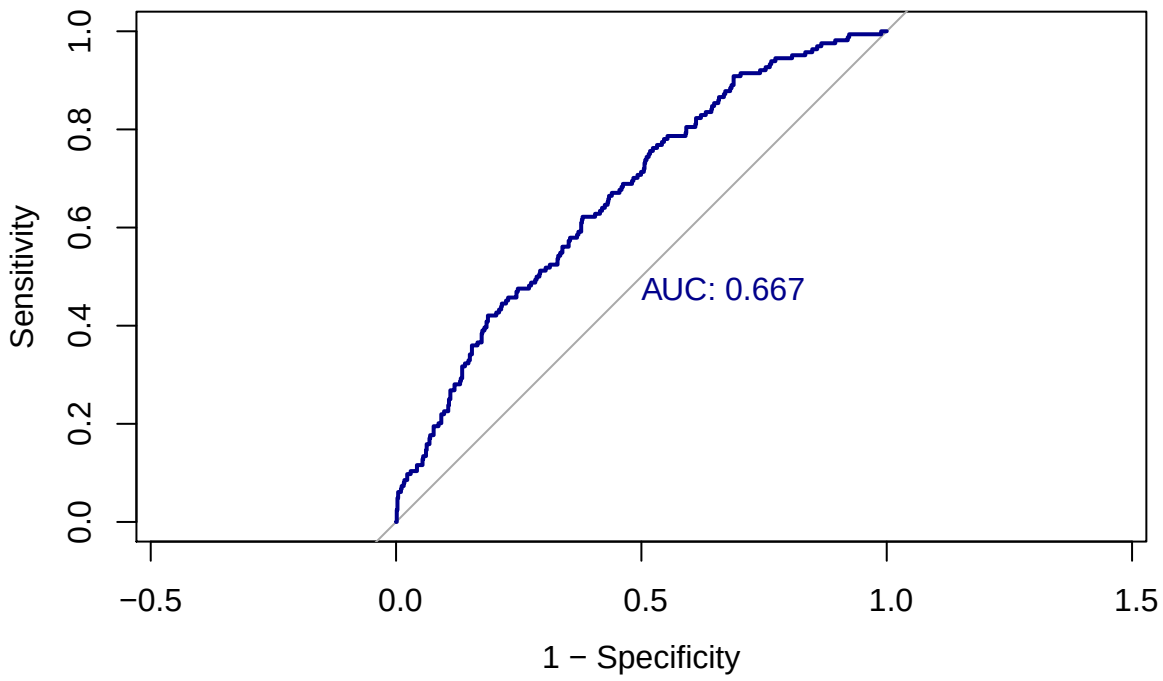
The confusion matrix is:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 437  62
##          1 268 102
##
##                Accuracy : 0.6203
##                  95% CI : (0.587, 0.6526)
##     No Information Rate : 0.8113
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1632
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6220
##             Specificity : 0.6199
##          Pos Pred Value : 0.2757
##          Neg Pred Value : 0.8758
##              Prevalence : 0.1887
##          Detection Rate : 0.1174
##    Detection Prevalence : 0.4258
##       Balanced Accuracy : 0.6209
##
##        'Positive' Class : 1
##
```

Roc curve is:

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

## Setting direction: controls < cases
```



Model 2 (mage_c + mht_c + mpregwt_c + med + mrace + smoke)- summary

The model summary is:

```
##
## Call:
## glm(formula = Premature ~ mage_c + mht_c + mpregwt_c + med +
##     mrace + smoke, family = binomial, data = smoking_filtered)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7305  -0.6719  -0.5600  -0.4104   2.4259
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.037096   0.959934  -1.080 0.279972
## mage_c       0.013191   0.016293   0.810 0.418185
## mht_c       -0.025305   0.042102  -0.601 0.547809
## mpregwt_c   -0.011334   0.005517  -2.054 0.039932 *
## med1        -0.368140   0.970391  -0.379 0.704411
## med2        -0.729035   0.959596  -0.760 0.447415
## med3        -0.584686   1.005457  -0.582 0.560895
## med4        -1.367352   0.976769  -1.400 0.161551
```

```
## med5        -0.916740   0.974744  -0.940 0.346965
## med7         1.941310   1.491497   1.302 0.193058
## mrace6       0.111383   0.523787   0.213 0.831601
## mrace7       0.748429   0.225459   3.320 0.000902 ***
## mrace8       0.835251   0.414498   2.015 0.043895 *
## mrace9      -0.779766   1.053791  -0.740 0.459323
## smoke1       0.301436   0.184815   1.631 0.102887
## ---
## 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: 794.80  on 854  degrees of freedom
## AIC: 824.8
##
## Number of Fisher Scoring iterations: 5
```

On doing a stepwise analysis (using AIC) of this model, the results are:

```
##
## Call:  glm(formula = Premature ~ med + mrace + mpregwt_c + smoke, family = binomial,
##     data = smoking_filtered)
##
## Coefficients:
## (Intercept)         med1         med2         med3         med4
##    -0.87665     -0.54103     -0.88762     -0.70667     -1.54789
##        med5         med7       mrace6       mrace7       mrace8
##    -1.06462      1.82568      0.15486      0.77069      0.90597
##      mrace9    mpregwt_c       smoke1
##    -0.75284     -0.01215      0.28889
##
## Degrees of Freedom: 868 Total (i.e. Null);  856 Residual
## Null Deviance:       841.8
## Residual Deviance: 795.9     AIC: 821.9
```

The confusion matrix is:

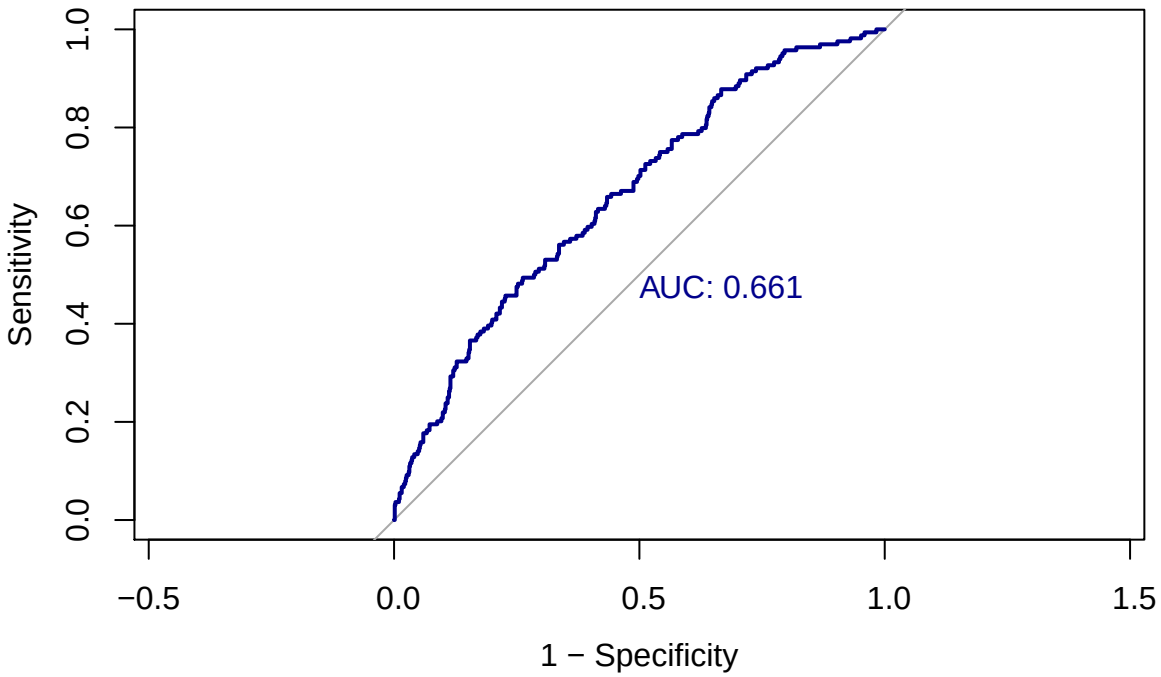
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 441  69
##          1 264  95
##
##                Accuracy : 0.6168
##                  95% CI : (0.5835, 0.6493)
##     No Information Rate : 0.8113
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1406
##
##  Mcnemar's Test P-Value : <2e-16
```

```
##
##             Sensitivity : 0.5793
##             Specificity : 0.6255
##          Pos Pred Value : 0.2646
##          Neg Pred Value : 0.8647
##              Prevalence : 0.1887
##          Detection Rate : 0.1093
##    Detection Prevalence : 0.4131
##       Balanced Accuracy : 0.6024
##
##        'Positive' Class : 1
##
```

Roc curve is:

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

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



Model 3 (mpregwt_c + med + mrace + smoke)- summary

The summary of this model is:

```
##
## Call:
## glm(formula = Premature ~ mpregwt_c + med + mrace + smoke, family = binomial,
##     data = smoking_filtered)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7356  -0.6722  -0.5576  -0.4106   2.4417
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.876651   0.940342  -0.932 0.351198
## mpregwt_c   -0.012147   0.004844  -2.508 0.012147 *
## med1        -0.541029   0.948966  -0.570 0.568593
## med2        -0.887619   0.940660  -0.944 0.345368
## med3        -0.706669   0.993071  -0.712 0.476713
## med4        -1.547889   0.955839  -1.619 0.105360
## med5        -1.064617   0.958531  -1.111 0.266708
## med7         1.825677   1.484105   1.230 0.218640
## mrace6       0.154865   0.516398   0.300 0.764258
## mrace7       0.770692   0.222714   3.460 0.000539 ***
## mrace8       0.905970   0.407690   2.222 0.026269 *
## mrace9      -0.752839   1.051521  -0.716 0.474020
## smoke1       0.288894   0.184335   1.567 0.117062
## ---
## 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: 795.91  on 856  degrees of freedom
## AIC: 821.91
##
## Number of Fisher Scoring iterations: 5
```

The confusion matrix of this model is:

```
#let's do the confusion matrix
Conf_mat3 <- confusionMatrix(as.factor(ifelse(fitted(smokreg3) >= reference_threshold, "1","0")),
                             as.factor(smoking_filtered$Premature),positive = "1")
Conf_mat3
```

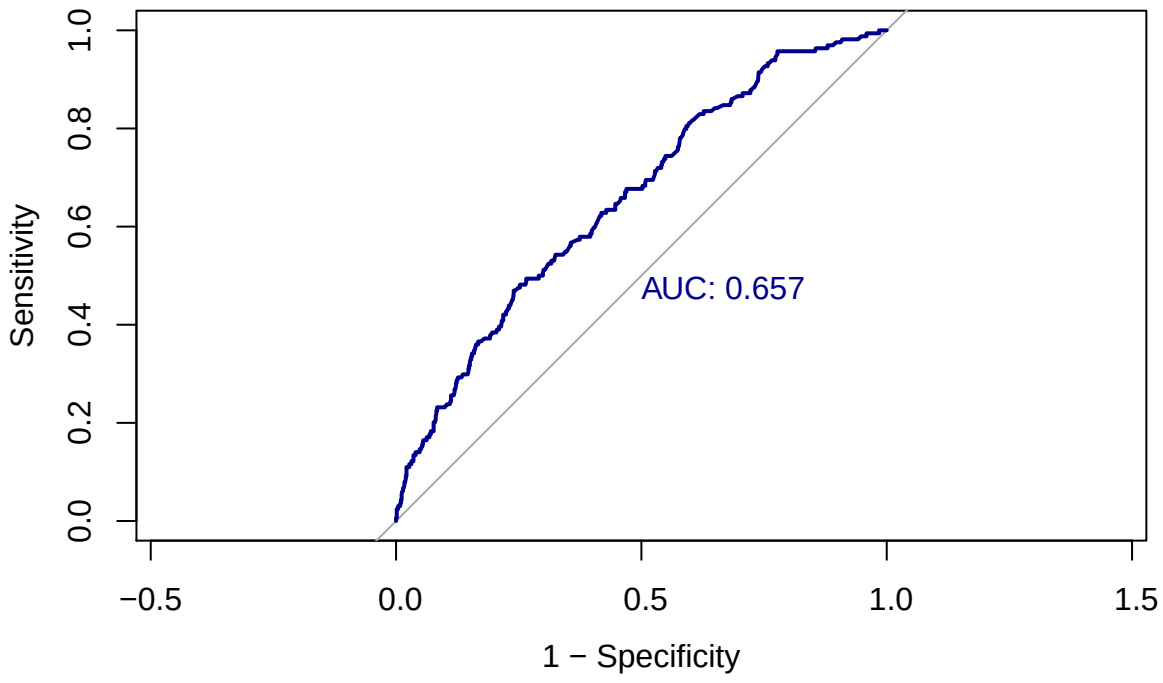
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 434  69
##          1 271  95
##
##                Accuracy : 0.6087
##                  95% CI : (0.5754, 0.6414)
##     No Information Rate : 0.8113
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1323
##
##  Mcnemar's Test P-Value : <2e-16
```

```
##
##             Sensitivity : 0.5793
##             Specificity : 0.6156
##          Pos Pred Value : 0.2596
##          Neg Pred Value : 0.8628
##              Prevalence : 0.1887
##          Detection Rate : 0.1093
##    Detection Prevalence : 0.4212
##       Balanced Accuracy : 0.5974
##
##        'Positive' Class : 1
##
```

Roc curve is:

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

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



Model 4 (mpregwt_c + med + mrace + smoke + mrace:smoke + smoke:mpregwt_c + mrace:mpregwt_c)- summary

The model summary is:

```
smokreg4 = glm(Premature ~ mpregwt_c + med + mrace + smoke + mrace*smoke + smoke*mpregwt_c + mrace*mpreg
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(smokreg4)
```

```
##
## Call:
## glm(formula = Premature ~ mpregwt_c + med + mrace + smoke + mrace *
##     smoke + smoke * mpregwt_c + mrace * mpregwt_c, family = binomial,
##     data = smoking_filtered)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7284  -0.6718  -0.5461  -0.3917   2.3793
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -8.524e-01  9.594e-01  -0.888 0.374301
## mpregwt_c        -5.729e-03  8.550e-03  -0.670 0.502795
## med1             -5.749e-01  9.655e-01  -0.595 0.551536
## med2             -9.549e-01  9.624e-01  -0.992 0.321082
## med3             -8.055e-01  1.017e+00  -0.792 0.428439
## med4             -1.680e+00  9.782e-01  -1.717 0.085984 .
## med5             -1.116e+00  9.790e-01  -1.140 0.254124
## med7              1.777e+00  1.507e+00   1.179 0.238328
## mrace6            2.003e-01  6.931e-01   0.289 0.772631
## mrace7            1.072e+00  3.121e-01   3.434 0.000595 ***
## mrace8            1.161e+00  7.989e-01   1.453 0.146225
## mrace9           -6.223e+01  2.423e+03  -0.026 0.979507
## smoke1            3.836e-01  2.297e-01   1.670 0.094954 .
## mrace6:smoke1    -1.149e-01  1.137e+00  -0.101 0.919494
## mrace7:smoke1    -5.720e-01  4.358e-01  -1.313 0.189349
## mrace8:smoke1     2.503e-01  8.634e-01   0.290 0.771907
## mrace9:smoke1     4.383e+01  1.858e+03   0.024 0.981177
## mpregwt_c:smoke1 -5.392e-03  9.924e-03  -0.543 0.586883
## mpregwt_c:mrace6 -3.417e-03  3.183e-02  -0.107 0.914517
## mpregwt_c:mrace7 -1.058e-02  1.063e-02  -0.995 0.319526
## mpregwt_c:mrace8  1.115e-02  3.441e-02   0.324 0.745931
## mpregwt_c:mrace9 -1.371e+00  5.931e+01  -0.023 0.981554
## ---
## 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: 785.14  on 847  degrees of freedom
## AIC: 829.14
##
## Number of Fisher Scoring iterations: 17
```

The step wise analysis (using AIC) results of this model gives:

```
##
## Call:  glm(formula = Premature ~ med + mrace + mpregwt_c + smoke, family = binomial,
##     data = smoking_filtered)
##
```

```
## Coefficients:
## (Intercept)         med1         med2         med3         med4
##    -0.87665     -0.54103     -0.88762     -0.70667     -1.54789
##        med5         med7       mrace6       mrace7       mrace8
##    -1.06462      1.82568      0.15486      0.77069      0.90597
##      mrace9    mpregwt_c       smoke1
##    -0.75284     -0.01215      0.28889
##
## Degrees of Freedom: 868 Total (i.e. Null);  856 Residual
## Null Deviance:       841.8
## Residual Deviance: 795.9     AIC: 821.9
```

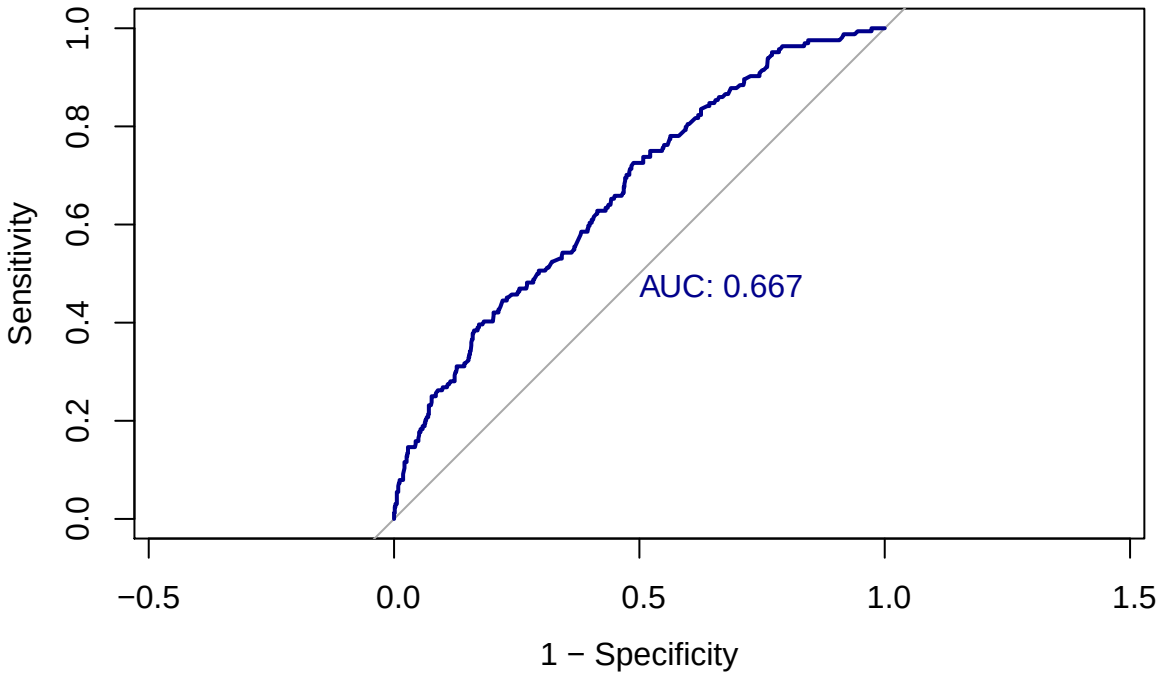
The confusion matrix is:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 427  68
##          1 278  96
##
##                Accuracy : 0.6018
##                  95% CI : (0.5684, 0.6346)
##     No Information Rate : 0.8113
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1281
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.5854
##             Specificity : 0.6057
##          Pos Pred Value : 0.2567
##          Neg Pred Value : 0.8626
##              Prevalence : 0.1887
##          Detection Rate : 0.1105
##    Detection Prevalence : 0.4304
##       Balanced Accuracy : 0.5955
##
##        'Positive' Class : 1
##
```

Roc curve is:

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

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



Model 5 (mpregwt_c + med + mrace + smoke + mrace:smoke)- summary

Model summary is:

```
##
## Call:
## glm(formula = Premature ~ mpregwt_c + med + mrace + smoke + mrace:smoke,
##     family = binomial, data = smoking_filtered)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7382  -0.6788  -0.5491  -0.3987   2.4788
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.923688   0.956831  -0.965 0.334364
## mpregwt_c      -0.012685   0.004833  -2.625 0.008678 **
## med1           -0.559196   0.963888  -0.580 0.561816
## med2           -0.906427   0.960176  -0.944 0.345158
## med3           -0.741215   1.014401  -0.731 0.464967
## med4           -1.574408   0.973899  -1.617 0.105964
## med5           -1.063172   0.976864  -1.088 0.276440
## med7            1.839205   1.506157   1.221 0.222039
## mrace6          0.187391   0.629208   0.298 0.765840
## mrace7          1.055195   0.305807   3.451 0.000559 ***
## mrace8          0.827330   0.494692   1.672 0.094442 .
## mrace9        -13.514954 413.950462  -0.033 0.973955
## smoke1          0.397100   0.227898   1.742 0.081430 .
```

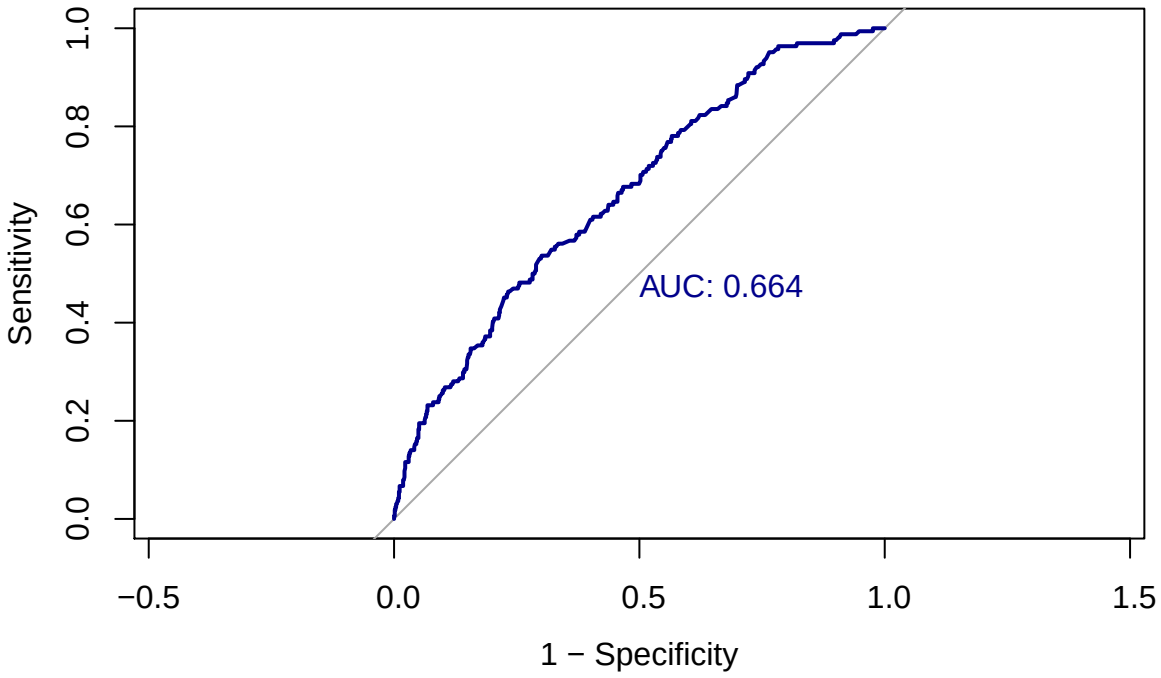
```
## mrace6:smoke1  -0.032475   1.112534  -0.029 0.976713
## mrace7:smoke1  -0.565223   0.424075  -1.333 0.182585
## mrace8:smoke1   0.317030   0.845121   0.375 0.707564
## mrace9:smoke1  14.462424 413.952402   0.035 0.972130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 841.83  on 868  degrees of freedom
## Residual deviance: 790.73  on 852  degrees of freedom
## AIC: 824.73
##
## Number of Fisher Scoring iterations: 14
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 423  64
##          1 282 100
##
##                Accuracy : 0.6018
##                  95% CI : (0.5684, 0.6346)
##     No Information Rate : 0.8113
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1389
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6098
##             Specificity : 0.6000
##          Pos Pred Value : 0.2618
##          Neg Pred Value : 0.8686
##              Prevalence : 0.1887
##          Detection Rate : 0.1151
##    Detection Prevalence : 0.4396
##       Balanced Accuracy : 0.6049
##
##        'Positive' Class : 1
##
```

The ROC curve is as follows:

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

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



ANOVA tests for comparing model 3 and model 5

```
anova(smokreg5, smokreg3, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ mpregwt_c + med + mrace + smoke + mrace:smoke
## Model 2: Premature ~ mpregwt_c + med + mrace + smoke
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       852     790.73
## 2       856     795.91 -4  -5.1847   0.2689
```

Binned Residual Plot for model 5