IDS 702 - Homework #4

Ana Belen Barcenas J.

10/10/2018

Maternal Smoking and Pre-term Birth

Let's create the dataset:

```
smoking_comp <- subset(smoking_comp, select = c(id, Premature))
smoking2 <- merge(smoking, smoking_comp, by="id")
```

Let's take a look of the data we have:

```
summary(smoking2)
```

```
##        id            date        gestation         bwt.oz     
##  Min.   :  15   Min.   :1350   Min.   :148.0   Min.   : 55.0  
##  1st Qu.:5477   1st Qu.:1444   1st Qu.:272.0   1st Qu.:108.0  
##  Median :6734   Median :1540   Median :279.0   Median :119.0  
##  Mean   :6032   Mean   :1536   Mean   :278.5   Mean   :118.4  
##  3rd Qu.:7587   3rd Qu.:1627   3rd Qu.:286.0   3rd Qu.:129.0  
##  Max.   :9263   Max.   :1714   Max.   :338.0   Max.   :174.0  
##      parity           mrace            mage            med       
##  Min.   : 0.000   Min.   :0.000   Min.   :15.00   Min.   :0.000  
##  1st Qu.: 1.000   1st Qu.:0.000   1st Qu.:23.00   1st Qu.:2.000  
##  Median : 2.000   Median :2.000   Median :26.00   Median :2.000  
##  Mean   : 1.953   Mean   :2.995   Mean   :27.29   Mean   :2.932  
##  3rd Qu.: 3.000   3rd Qu.:7.000   3rd Qu.:31.00   3rd Qu.:4.000  
##  Max.   :11.000   Max.   :9.000   Max.   :45.00   Max.   :7.000  
##       mht           mpregwt           inc            smoke       
##  Min.   :53.00   Min.   : 87.0   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:62.00   1st Qu.:113.0   1st Qu.:2.000   1st Qu.:0.0000  
##  Median :64.00   Median :125.0   Median :3.000   Median :0.0000  
##  Mean   :64.07   Mean   :128.5   Mean   :3.681   Mean   :0.4638  
##  3rd Qu.:66.00   3rd Qu.:140.0   3rd Qu.:5.000   3rd Qu.:1.0000  
##  Max.   :72.00   Max.   :220.0   Max.   :9.000   Max.   :1.0000  
##    Premature     
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0000  
##  Mean   :0.1887  
##  3rd Qu.:0.0000  
##  Max.   :1.0000
```

The complete dataset includes father's data as well as mother's data. In the previous assignment (Methods and Data Analysis #3) I found out that father's data has a strong correlation with mother's data. Moreover, after deleating missing values in father's information I lose almost half of the observations. Thus, I will not use father's information to predict the effect of maternal smoking on pre-term birth.

And the predictor variable:

```
smoking2 %>% count(Premature)
```

```
## # A tibble: 2 x 2
##   Premature     n
##       <int> <int>
## 1         0   705
## 2         1   164
```

It seems that the databased is not balanced in terms of premature babies. We have only 19% premature babies born. We will work with the data we have having in mind that this is a possible limitation of the analysis.

Creation of possible predictors

1) Collapsing race categories from 0 to 5 in "white" and create dummy vars for each category:

```
n = nrow(smoking2)
smoking2$white = rep(0, n)
smoking2$white[smoking2$mrace == "0" | smoking2$mrace == "1" | smoking2$mrace == "2" | smoking2$mrace =
smoking2$mexican = rep(0, n)
smoking2$mexican[smoking2$mrace == "6"] = 1
smoking2$black = rep(0, n)
smoking2$black[smoking2$mrace == "7"] = 1
smoking2$asian = rep(0, n)
smoking2$asian[smoking2$mrace == "8"] = 1
smoking2$mix = rep(0, n)
smoking2$mix[smoking2$mrace == "9"] = 1

smoking2$mrace2[smoking2$mrace == "0" | smoking2$mrace == "1" | smoking2$mrace == "2" | smoking2$mrace
smoking2$mrace2[smoking2$mrace == "6"] = "Mexican"
smoking2$mrace2[smoking2$mrace == "7"] = "Black"
smoking2$mrace2[smoking2$mrace == "8"] = "Asian"
smoking2$mrace2[smoking2$mrace == "9"] = "Mix"

smoking2 %>% count(mrace2)
```

```
## # A tibble: 5 x 2
##   mrace2      n
##   <chr>   <int>
## 1 Asian      34
## 2 Black     169
## 3 Mexican    25
## 4 Mix        15
## 5 white     626
```

The number of observations we have in the categories different than "white" is really small compared with the "white" mothers. I would perform a second transformation of race: white and other race:

```
smoking2$white_dum[smoking2$mrace == "0" | smoking2$mrace == "1" | smoking2$mrace == "2" | smoking2$mra

smoking2$white_dum[smoking2$mrace == "6" | smoking2$mrace == "7"| smoking2$mrace == "8" | smoking2$mrac

smoking2 %>% count(white_dum)
```

```
## # A tibble: 2 x 2
##   white_dum     n
##       <dbl> <int>
## 1      0      243
```

```
## 2      1.00   626
```

2) Let's analyze mother's education observations:

```r
smoking2 %>% count(med)
```

```
## # A tibble: 7 x 2
##     med     n
##   <int> <int>
## 1     0     5
## 2     1   130
## 3     2   321
## 4     3    47
## 5     4   203
## 6     5   159
## 7     7     4
```

It seems that could be useful to collapse the categories in 4 clearer batches: less than high school education, high school education (and no other schooling), more than high school education but less than college, and college graduate.

```r
n = nrow(smoking2)
smoking2$less_hs = rep(0, n)
smoking2$less_hs[smoking2$med == "0" | smoking2$med == "1"] = 1
smoking2$hs = rep(0, n)
smoking2$hs[smoking2$med == "2"] = 1
smoking2$more_hs = rep(0, n)
smoking2$more_hs[smoking2$med == "3" | smoking2$med == "4" | smoking2$med == "6" | smoking2$med == "7"]
smoking2$col = rep(0, n)
smoking2$col[smoking2$med == "5"] = 1

smoking2$med2[smoking2$med == "0" | smoking2$med == "1"] = "less_hs"
smoking2$med2[smoking2$med == "2"] = "hs"
smoking2$med2[smoking2$med == "3" | smoking2$med == "4" | smoking2$med == "6" | smoking2$med == "7"] =
smoking2$med2[smoking2$med == "5"] = "college"

smoking2 %>% count(med2)
```

```
## # A tibble: 4 x 2
##   med2        n
##   <chr>   <int>
## 1 college   159
## 2 hs        321
## 3 less_hs   135
## 4 more_hs   254
```

3) To obtain more accurate interpretations, I'll substract the mean of the mother's age, height and weight:

```r
smoking2$mage_cent = smoking2$mage - mean(smoking2$mage)
smoking2$mht_cent = smoking2$mht - mean(smoking2$mht)
smoking2$mpregwt_cent = smoking2$mpregwt - mean(smoking2$mpregwt)
```

Exploratory analysis

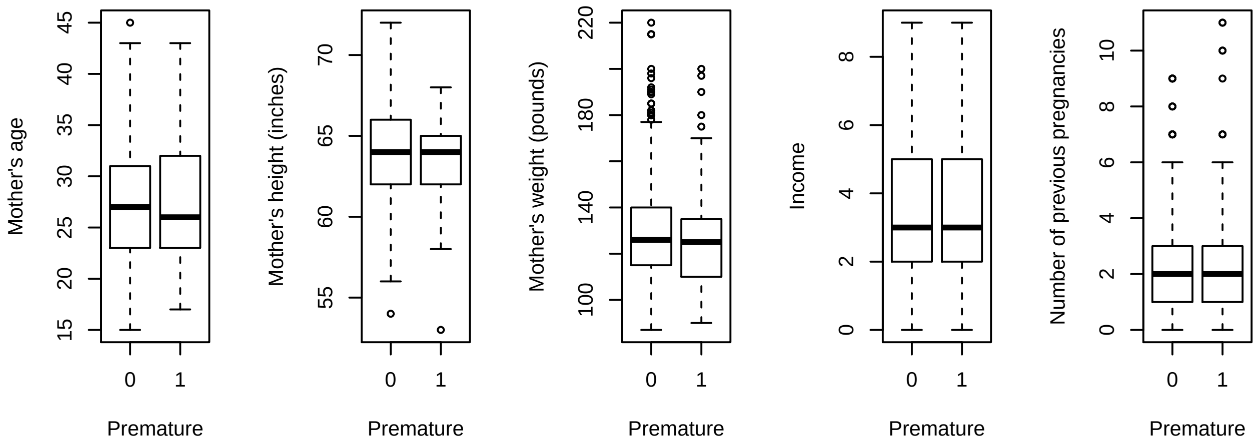
The predictor variables available are: – Mother's race or ethnicity – Mother's age – Mother's education – Mother's height – Mother's pre pregnancy weight – Income – Parity (number of previous pregnancies) –

Smoke

Let's analyze the individual effect that the predictors has on pre-term births as well as the combined effect of smoking, race and pre-term births:

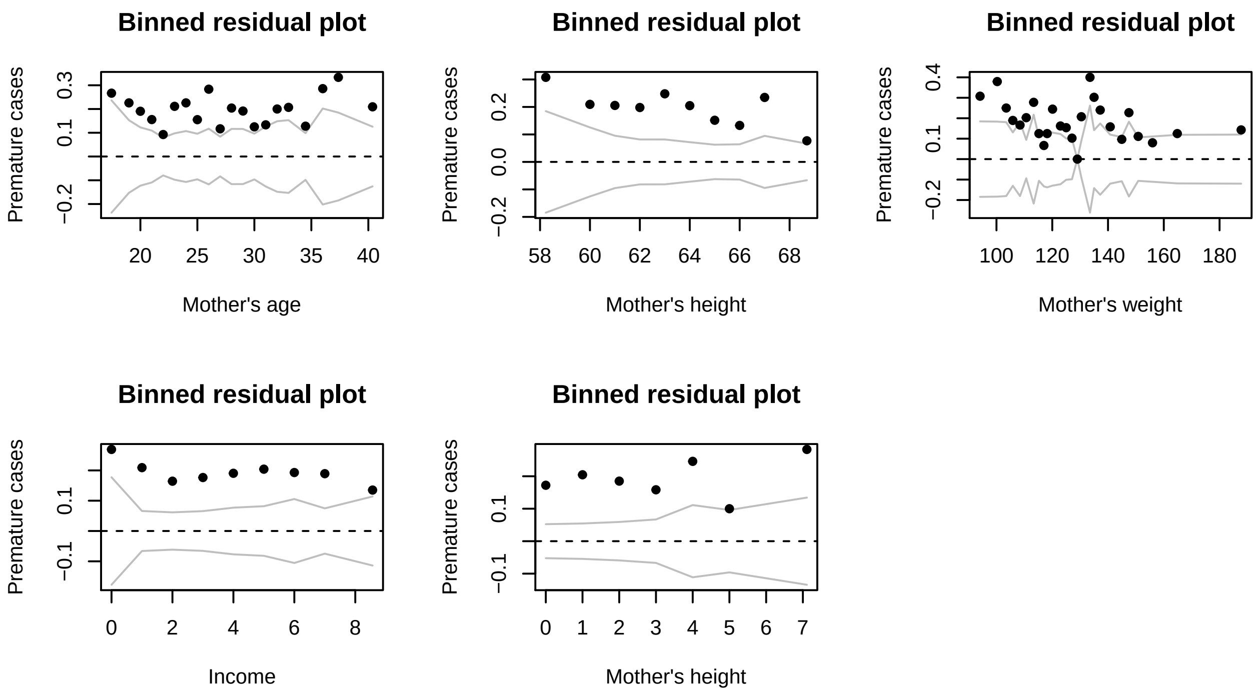
- Continous predictors vs. pre-term birth:

```
par(mfrow=c(1,5))
boxplot(smoking2$mage~smoking2$Premature, ylab = "Mother's age", xlab = "Premature")
boxplot(smoking2$mht~smoking2$Premature, ylab = "Mother's height (inches)", xlab = "Premature")
boxplot(smoking2$mpregwt~smoking2$Premature, ylab = "Mother's weight (pounds)", xlab = "Premature")
boxplot(smoking2$inc~smoking2$Premature, ylab = "Income", xlab = "Premature")
boxplot(smoking2$parity~smoking2$Premature, ylab = "Number of previous pregnancies", xlab = "Premature"
```



From this boxplots is not evident that mother's height, weight, parity, and income can predict pre-term birth. Nonetheless, this relationships are individual. A logistic model should give us a better and more complete understanding of the relation between these variables and pre-term birth. Thus, we will test the predictive power of these predictors in the model. In the meantime I will look at the binned plots of these predictors vs premature:

```
par(mfrow=c(2,3))
binnedplot(smoking2$mage, y=smoking2$Premature, xlab = "Mother's age", ylab = "Premature cases")
binnedplot(smoking2$mht, y=smoking2$Premature, xlab = "Mother's height", ylab = "Premature cases")
binnedplot(smoking2$mpregwt, y=smoking2$Premature, xlab = "Mother's weight", ylab = "Premature cases")
binnedplot(smoking2$inc, y=smoking2$Premature, xlab = "Income", ylab = "Premature cases")
binnedplot(smoking2$parity, y=smoking2$Premature, xlab = "Mother's height", ylab = "Premature cases")
```



No transformations suggested based on the binned plots.

- Interactions with smoking and categorial predictors (education and race):

```
table(smoking2$smoke, smoking2$Premature)
```

```
##
##       0   1
##   0 389  77
##   1 316  87
```

```
table(smoking2$med2, smoking2$Premature)
```

```
##
##              0   1
##   college 132  27
##   hs      260  61
##   less_hs  97  38
##   more_hs 216  38
```

```
table(smoking2$white_dum, smoking2$Premature)
```

```
##
##       0   1
##   0 180  63
##   1 525 101
```

From these tables we can see that (1) if the mother smoke, the number of pre-term births is greater than if the mother do not smoke (87 vs 77 obs.); (2) mother's graduated from college are the ones with less pre-term births, the relation with education as a continous variable is not clear enough; (3) if the mother is white, the number of pre-term births is greater than if the mother is not white (101 vs. 63 obs.). Once again, this relations are individual relations between predictors. A logistic model will give us a better understanding of the predictive power of each variable.

We have seen individual relations, let's fit a model.

Fitting a logistic regression

```r
premature1 = glm(Premature ~ as.factor(smoke) + white_dum + mage_cent + mht_cent + mpregwt_cent + inc + 
summary(premature1)
```

```
##
## Call:
## glm(formula = Premature ~ as.factor(smoke) + white_dum + mage_cent +
##     mht_cent + mpregwt_cent + inc + parity + hs + more_hs + col,
##     family = binomial, data = smoking2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2015  -0.6744  -0.5711  -0.4522   2.3156
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.737862   0.338766  -2.178  0.02940 *
## as.factor(smoke)1  0.335776   0.181764   1.847  0.06470 .
## white_dum         -0.626832   0.199145  -3.148  0.00165 **
## mage_cent          0.021710   0.019821   1.095  0.27338
## mht_cent          -0.033205   0.041311  -0.804  0.42152
## mpregwt_cent      -0.010174   0.005284  -1.926  0.05416 .
## inc                0.008374   0.042495   0.197  0.84378
## parity            -0.020559   0.058792  -0.350  0.72657
## hs                -0.417269   0.259478  -1.608  0.10781
## more_hs           -0.772625   0.285283  -2.708  0.00676 **
## col               -0.587395   0.330618  -1.777  0.07562 .
## ---
## 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: 810.07  on 858  degrees of freedom
## AIC: 832.07
##
## Number of Fisher Scoring iterations: 4
```

From the exploratory analysis we saw that the variables income and parity seemed to be non related with pre term birth. To conclude this properly, I will test how significant they are employing an anova:

```r
premature2 = glm(Premature ~ smoke + white_dum + mage_cent + mht_cent + mpregwt_cent + hs + more_hs + c

anova(premature1, premature2, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ as.factor(smoke) + white_dum + mage_cent + mht_cent +
##     mpregwt_cent + inc + parity + hs + more_hs + col
## Model 2: Premature ~ smoke + white_dum + mage_cent + mht_cent + mpregwt_cent +
##     hs + more_hs + col
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       858     810.07
## 2       860     810.24 -2  -0.1717   0.9177
```

Given the p-value we do not have enough information to reject the null hypotheses: it seems that the income and parity are not significant predictors of pre-term birth. I will exclude those variables.

Let's see if the relation between smoking and pre-term birth differs by mother's race:

```
premature3 = glm(Premature ~ smoke*white_dum + mage_cent + mht_cent + mpregwt_cent + hs + more_hs + col

anova(premature2, premature3, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Premature ~ smoke + white_dum + mage_cent + mht_cent + mpregwt_cent +
##     hs + more_hs + col
## Model 2: Premature ~ smoke * white_dum + mage_cent + mht_cent + mpregwt_cent +
##     hs + more_hs + col
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       860     810.24
## 2       859     809.68  1  0.55992   0.4543
```

It seems that the interaction between mother's race and smoking is not significant. Which is to say, the association between smoking and pre-term birth does not differ between white mothers and mothers from different races. I will not include the interaction.

So, the final model would be the following:

```
summary(premature2)
```

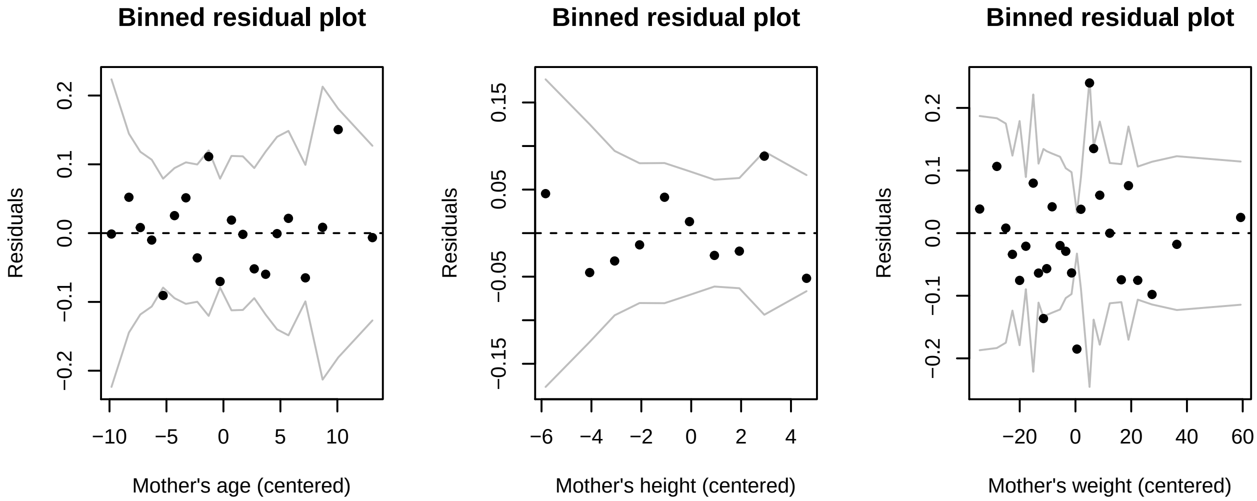
```
##
## Call:
## glm(formula = Premature ~ smoke + white_dum + mage_cent + mht_cent +
##     mpregwt_cent + hs + more_hs + col, family = binomial, data = smoking2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1921  -0.6762  -0.5709  -0.4520   2.3025
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.790845   0.247426  -3.196  0.00139 **
## smoke         0.335403   0.181415   1.849  0.06448 .
## white_dum    -0.611070   0.195408  -3.127  0.00177 **
## mage_cent     0.018767   0.015776   1.190  0.23420
## mht_cent     -0.032339   0.041185  -0.785  0.43233
## mpregwt_cent -0.010328   0.005268  -1.960  0.04996 *
## hs           -0.385429   0.247834  -1.555  0.11990
## more_hs      -0.735136   0.270281  -2.720  0.00653 **
## col          -0.531027   0.301314  -1.762  0.07801 .
## ---
## 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: 810.24  on 860  degrees of freedom
## AIC: 828.24
##
## Number of Fisher Scoring iterations: 4
```

Model diagnostics

Now that I have a model that seems to be good enough, let's check binned residuals, confussion matrix and ROC curve (predictions will com later after assessing the performance of the model).

1) Binned residual plots

```
par(mfrow=c(1,3))
rawresid = smoking2$Premature - fitted(premature2)
binnedplot(x=smoking2$mage_cent, y = rawresid, xlab = "Mother's age (centered)", ylab = "Residuals")
binnedplot(x=smoking2$mht_cent, y = rawresid, xlab = "Mother's height (centered)", ylab = "Residuals")
binnedplot(x=smoking2$mpregwt_cent, y = rawresid, xlab = "Mother's weight (centered)", ylab = "Residual
```



Residuals seems to be randomly distributed along the values and well distributed between positive and negative residuals. This suggests that the model is performing properly and describes the data good enough.

Let's look at average residuals by smoke, race and education:

```
tapply(rawresid, smoking2$smoke, mean)
```

```
##             0             1
## -8.033696e-12 -5.351221e-12
```

```
tapply(rawresid, smoking2$white_dum, mean)
```

```
##             0             1
## -7.153730e-12 -6.648384e-12
```

```
tapply(rawresid, smoking2$med2, mean)
```

```
##       college            hs       less_hs       more_hs
## -3.276474e-12 -3.807015e-12 -4.164008e-12 -1.415391e-11
```

Nothing remarkable from these residuals. Unless that the residual for mother's with less than high school education is big. This can be explained by the small number of observations we have for that category.

2) Lets see the confusion matrix with a 0.5 threshold:

```
threshold = 0.3
table(smoking2$Premature, premature2$fitted > threshold)
```

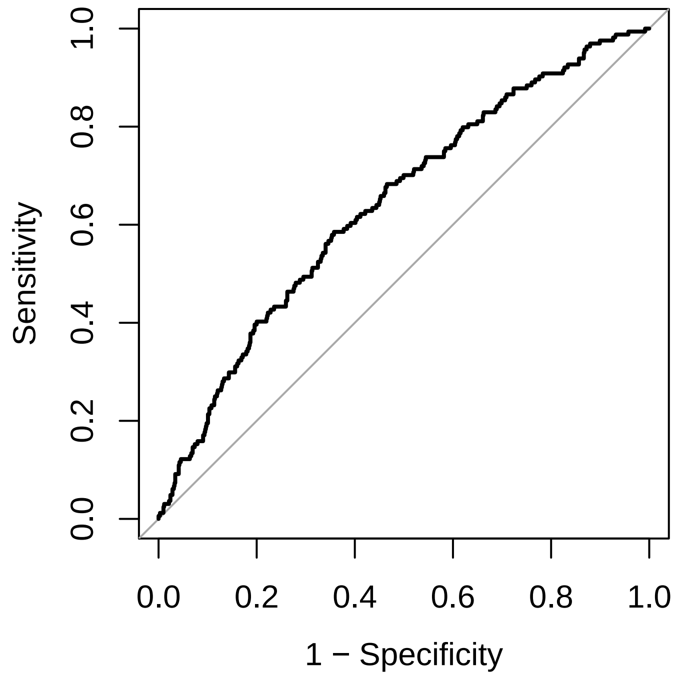
```
##
##     FALSE TRUE
##   0   650   55
```

```
##   1  139   25
```

It seems that the model is not doing so well in predicting true positives (premature births) but is performing better when predicting true negatives (when a mother will not have a premature birth). Let's check the ROC curve to determine if the model is better than a random choice. Also, do not forget that we do not have enough observations where the mother had a premature baby.

3) Finally, the ROC curve:

```
roc(smoking2$Premature, fitted(premature2), plot=T, legacy.axes=T)
```




```
##
## Call:
## roc.default(response = smoking2$Premature, predictor = fitted(premature2),     plot = T, legacy.axes
##
## Data: fitted(premature2) in 705 controls (smoking2$Premature 0) < 164 cases (smoking2$Premature 1).
## Area under the curve: 0.6402
```

The area under the curve is 0.6402, seems good! Seems to be a strong predictive logistics regression at least of true negative classifications.

Interpretations

```
exp(premature2$coefficients)
```

```
##  (Intercept)        smoke    white_dum    mage_cent     mht_cent
##    0.4534614    1.3985042    0.5427697    1.0189447    0.9681782
## mpregwt_cent           hs      more_hs          col
##    0.9897256    0.6801591    0.4794401    0.5880008
```

```
exp(confint.default(premature2))
```

```
##                 2.5 %    97.5 %
## (Intercept)  0.2792102 0.7364602
```

```
## smoke        0.9800378 1.9956514
## white_dum    0.3700698 0.7960632
## mage_cent    0.9879204 1.0509432
## mht_cent     0.8930962 1.0495724
## mpregwt_cent 0.9795585 0.9999982
## hs           0.4184602 1.1055205
## more_hs      0.2822740 0.8143249
## col          0.3257612 1.0613448
```

– Intercept: We expect the odds of having a premature baby given that the mother do not smoke, is not white, has average height, weight, and age, and has not high school degree, is 0.45. We are 95% confident that the odds when the person has the characteristics mentioned above falls between 0.28 and 0.74.

– Smoke: A mother with the characteristics mentioned above but that smokes, has a higher probability of having a premature baby. We expect the odds change by a multiplicative factor of 1.40 (CI: 0.98 - 2.00).

– Race: The odds that a mother that is white and has average height, weight, and age, and has not high school degree, will expect a change in the odds of having a premature baby by a multiplicative factor of 0.54 with respect to non-white mothers (CI:0.37 - 0.80).

– An interesting association found in the analysis, is that mother's weight seems to be a good predictor of pre-term birth. More specifically, A mother that is not white, has average height, weight, and age, and do not smoke will get an average change of the odds of having a premature baby by a multiplicative factor of 0.98 for each additional pound of weight (CI: 0.98 - 1.00).

Limitations of the analysis.

As mentioned above, I consider that the bigger limitation of this analysis is the amount of observations available. Specially those related with the mother's race. To determine if the relation between smoking and pre-term birth varies depending on the mother's race, I would perform an additional analysis including more information about mothers that are not white as well as more observations of mothers that indeed had a premature baby to get a more balanced dependent variable. Getting more information could also help us to predict more accurate true positives (predict that a mother will have a premature baby correctly).

Moreover, I would include a variable that describes the general health status of the mother before and during the pregnancy. Researchers has shown that mother's health is a strong predictor of new born babies health.