Homework4

Xuan Yu

10/05/2018

Maternal Smoking and Premature Birth

Read in the data.

```
library("pROC")
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library("arm")
```

```
## Loading required package: MASS
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.10-1, built: 2018-4-12)
```

```
## Working directory is /Users/xuanyu/Desktop/MIDS courses/data modeling/HW/HW4
```

```
maternal_data <- read.csv("/Users/xuanyu/Desktop/MIDS courses/data modeling/HW/HW4/smoking.csv")
maternal_data$premature <- rep(0, nrow(maternal_data))
maternal_data$premature[maternal_data$gestation < 270] <- 1

maternal_data$mrace_new <- maternal_data$mrace
maternal_data$mrace_new[maternal_data$mrace >= 0 & maternal_data$mrace <= 5] <- 5

maternal_data$who_smoke <- maternal_data$smoke
maternal_data$who_smoke[maternal_data$smoke != 0] <- 1

dim(maternal_data)
```

```
## [1] 869  15
```

```
summary(maternal_data)
```

```
##        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
```

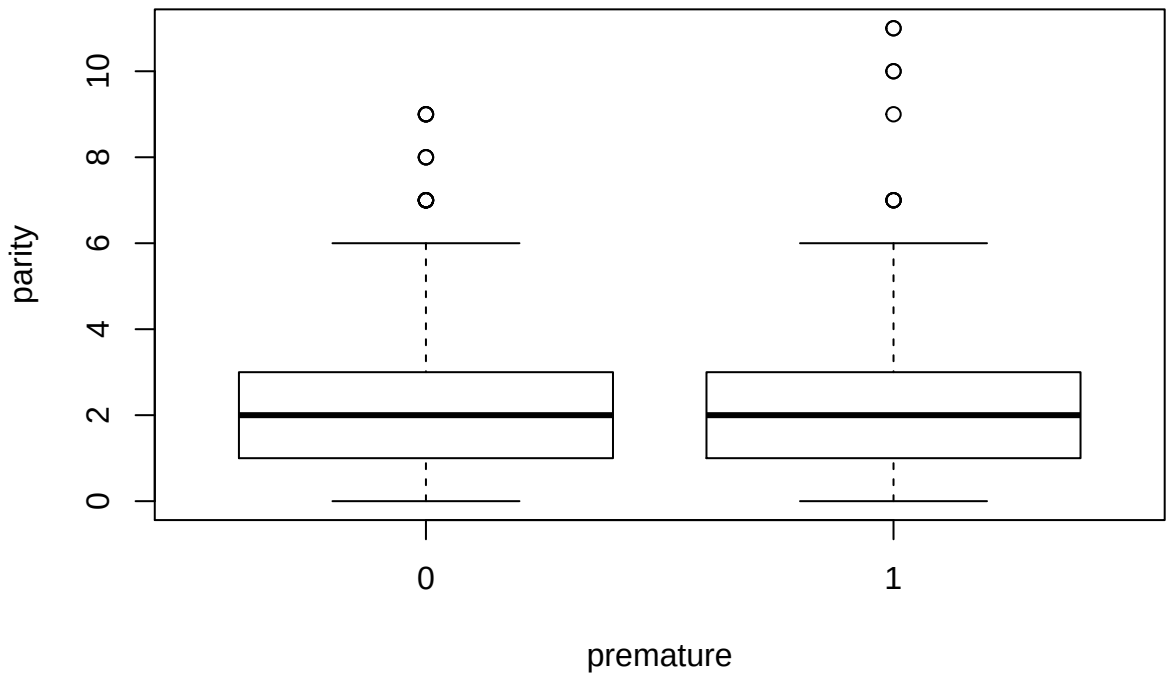
```
##  1st Qu.: 1.000   1st Qu.:0.000   1st Qu.:23.00   1st Qu.:2.000
##  Median : 2.000   Median :2.000   Median :26.00   Median :2.000
##  Mean   : 1.953   Mean   :2.995   Mean   :27.29   Mean   :2.932
##  3rd Qu.: 3.000   3rd Qu.:7.000   3rd Qu.:31.00   3rd Qu.:4.000
##  Max.   :11.000   Max.   :9.000   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        mrace_new       who_smoke
##  Min.   :0.0000   Min.   :5.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:5.000   1st Qu.:0.0000
##  Median :0.0000   Median :5.000   Median :0.0000
##  Mean   :0.1887   Mean   :5.604   Mean   :0.4638
##  3rd Qu.:0.0000   3rd Qu.:7.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :9.000   Max.   :1.0000
```

Some exploratory data analysis:

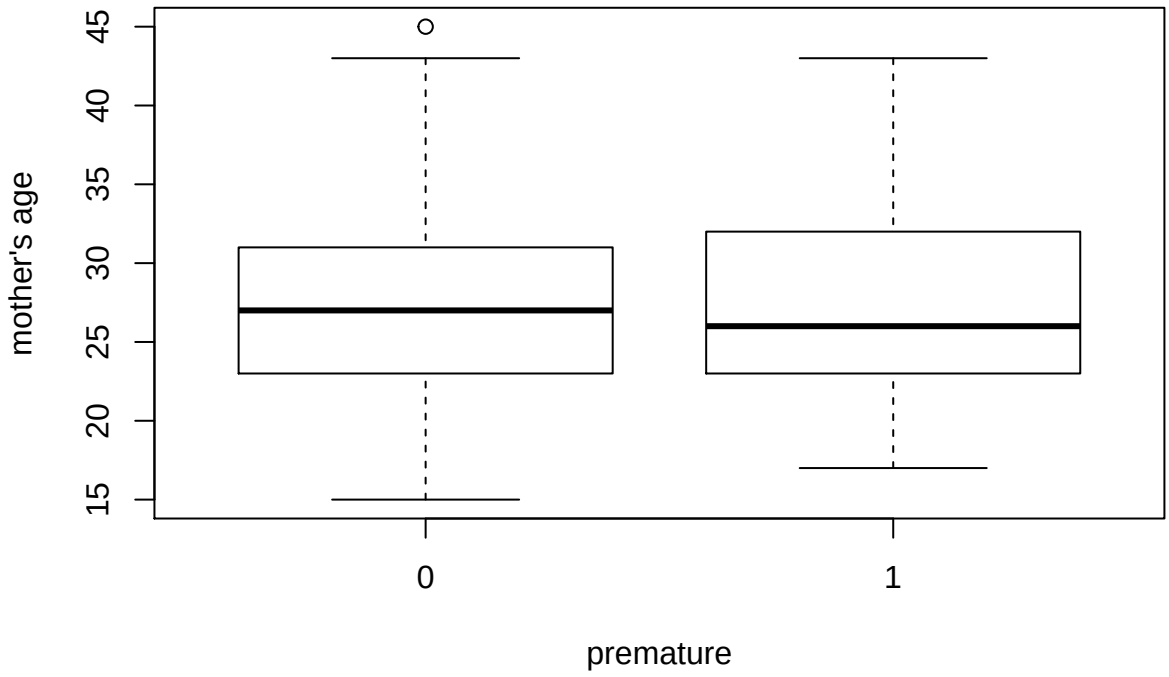
There are very few observation in education level 7 so we don't worried about its different mean value with other level

We may care more about smoke variable and race variable:

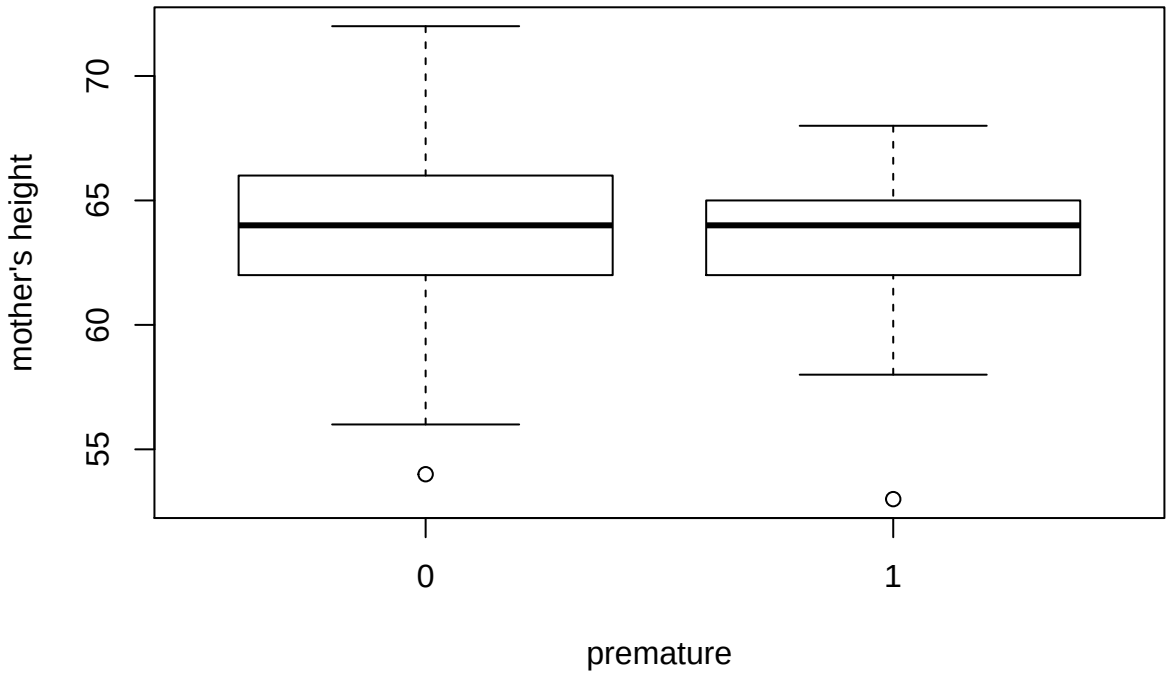
```
boxplot(parity~premature, data = maternal_data, xlab = "premature", ylab = "parity")
```



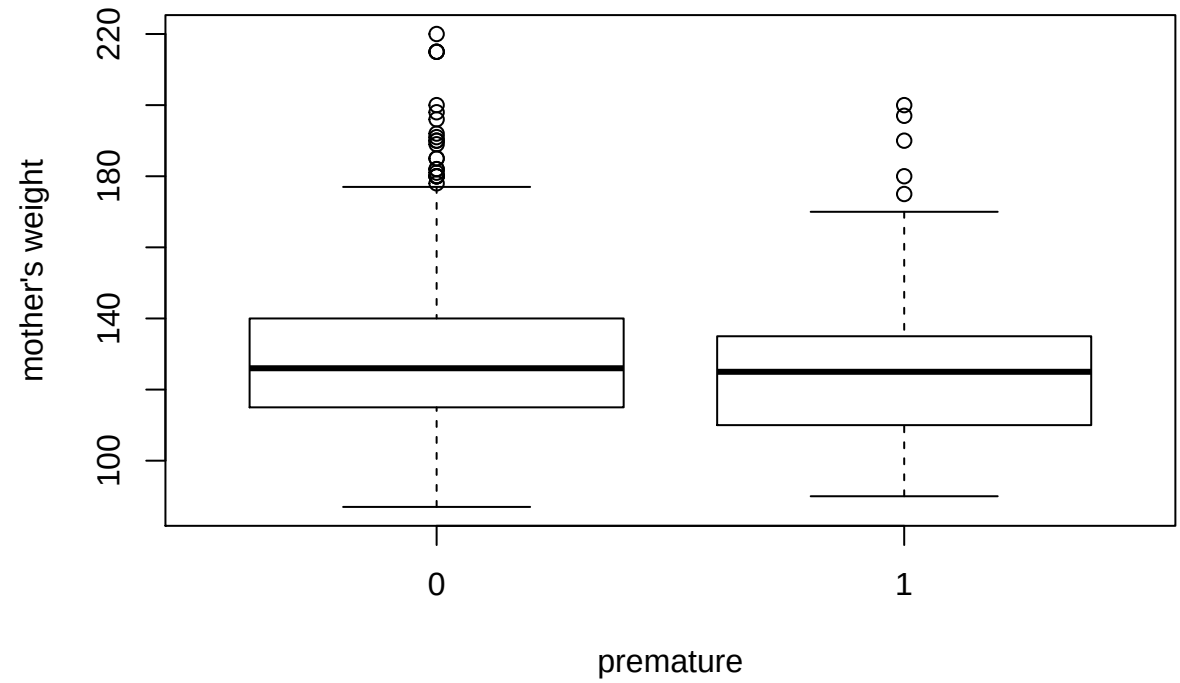

```
boxplot(mage~premature, data = maternal_data, xlab = "premature", ylab = "mother's age")
```



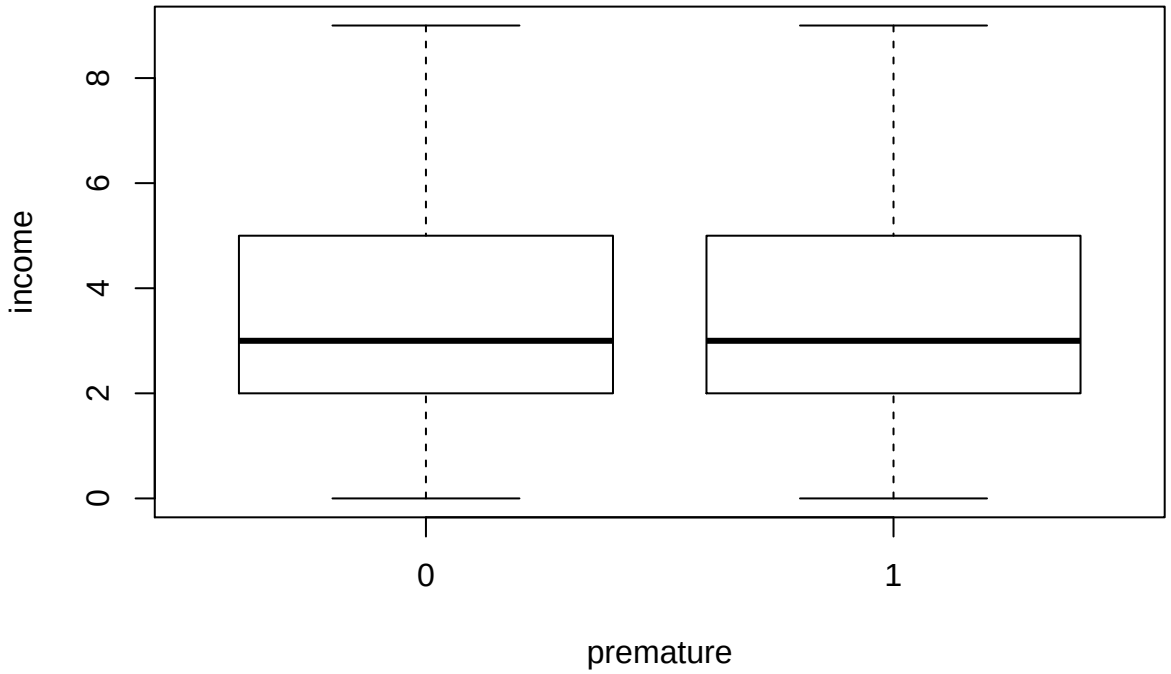

```
boxplot(mht~premature, data = maternal_data, xlab = "premature", ylab = "mother's height")
```




```
boxplot(mpregwt~premature, data = maternal_data, xlab = "premature", ylab = "mother's weight")
```




```
boxplot(inc~premature, data = maternal_data, xlab = "premature", ylab = "income")
```




```
tapply(maternal_data$premature, maternal_data$who_smoke, mean)
```

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

```
tapply(maternal_data$premature, maternal_data$mrace_new, mean)
```

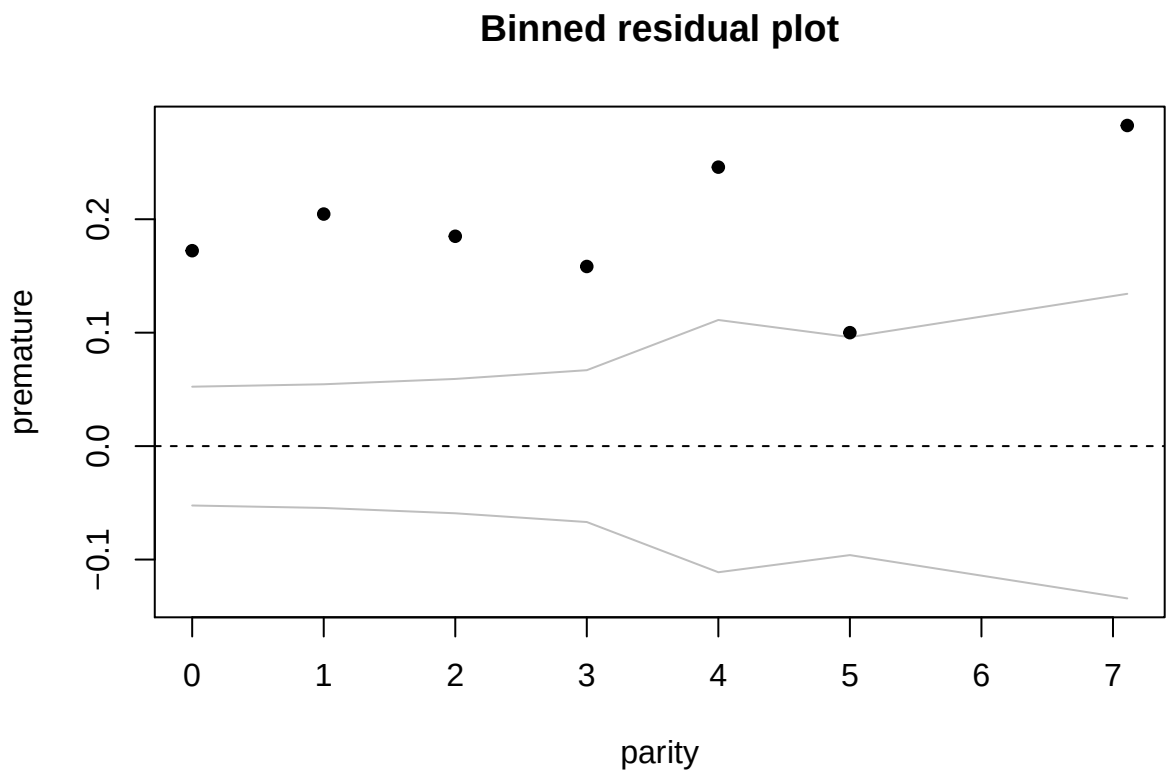
```
##          5          6          7          8          9
## 0.16134185 0.24000000 0.26627219 0.32352941 0.06666667
```

```
tapply(maternal_data$premature, maternal_data$med, mean)
```

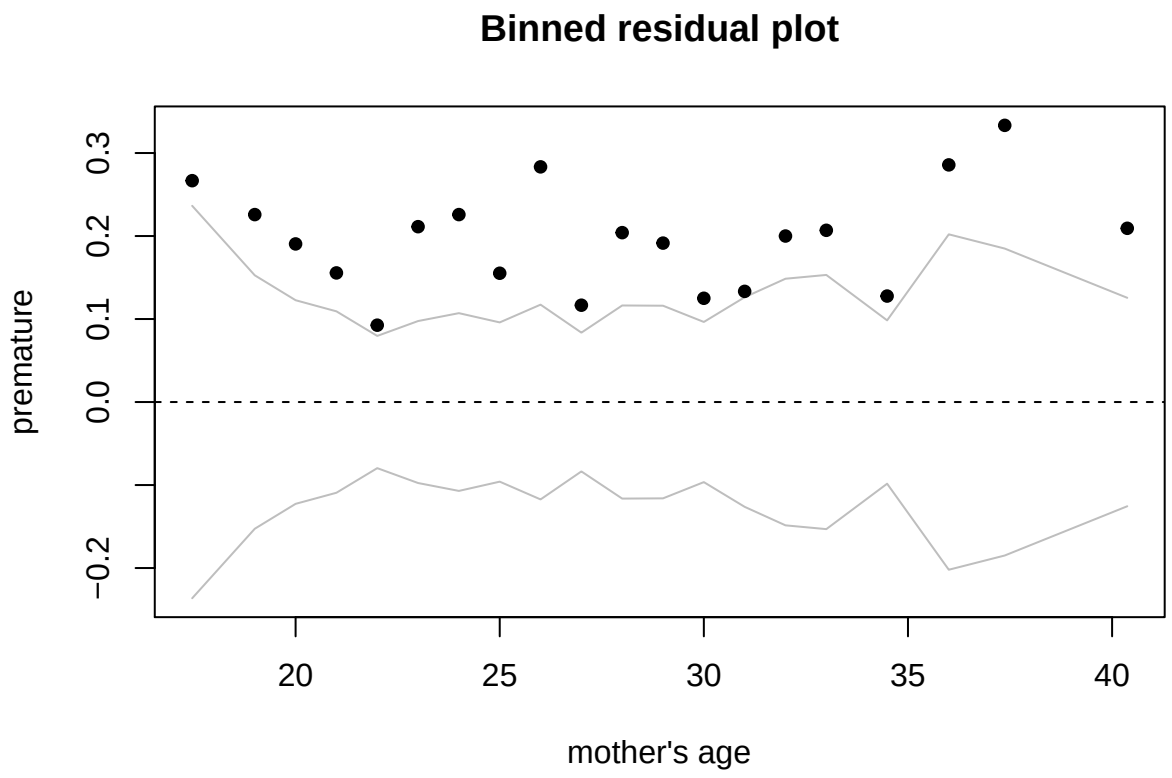
```
##         0         1         2         3         4         5         7
## 0.4000000 0.2769231 0.1900312 0.2340426 0.1182266 0.1698113 0.7500000
```

Now we are looking at binnedplots of continuous predictors versus premature birth: We'll ignore the SD lines in these plots – they are only relevant when plotting binned residuals versus the predicted probabilities:

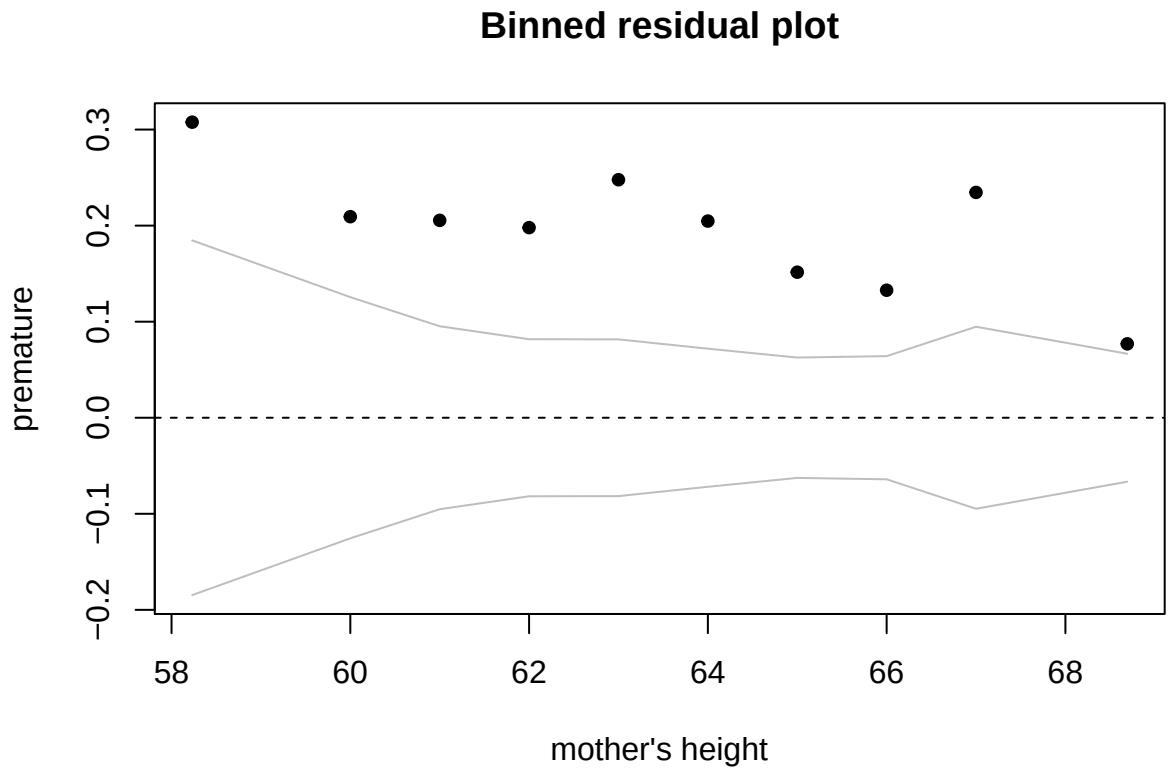
```
binnedplot(maternal_data$parity, y=maternal_data$premature, xlab = "parity", ylab = "premature")
```




```
binnedplot(maternal_data$mage, y=maternal_data$premature, xlab = "mother's age", ylab = "premature")
```

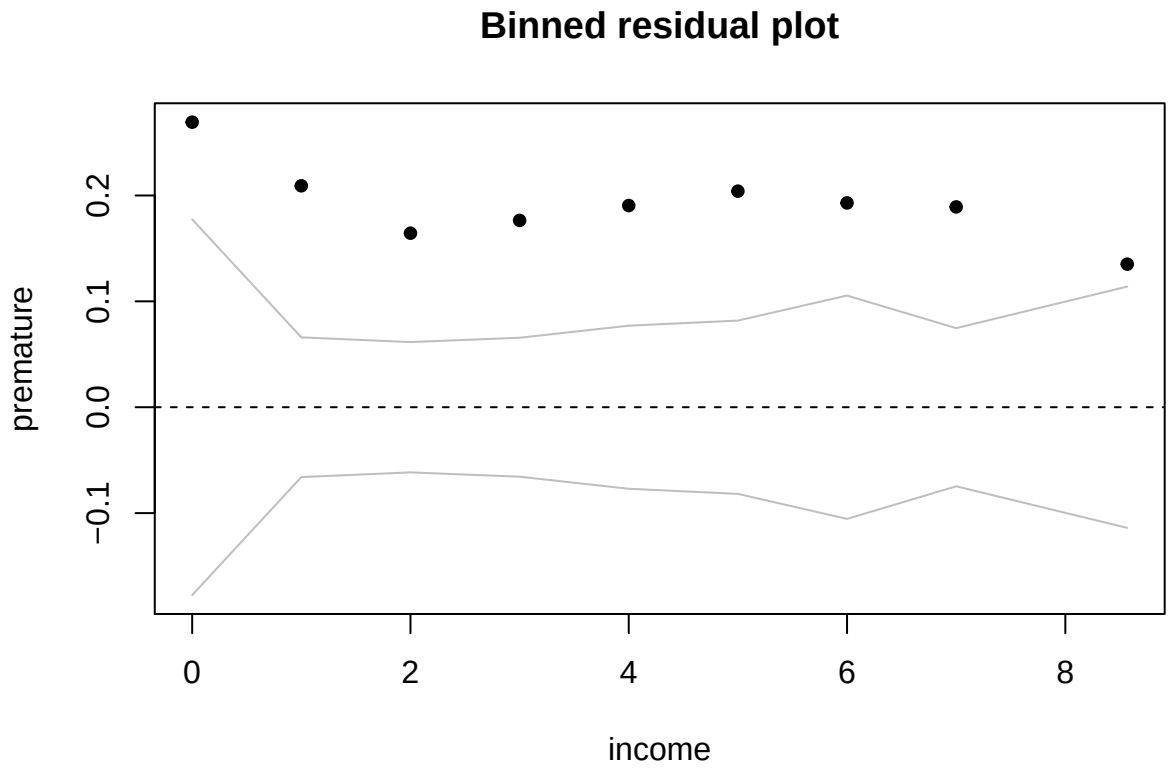



```
binnedplot(maternal_data$mht, y=maternal_data$premature, xlab = "mother's height", ylab = "premature")
```




```
binnedplot(maternal_data$mpregwt, y=maternal_data$premature, xlab = "mother's weight", ylab = "premature
```


```
binnedplot(maternal_data$inc, y=maternal_data$premature, xlab = "income", ylab = "premature")
```



Then try a logistic regression that has a main effect for every variable and linear predictors. Begin by centering the continuous predictors:

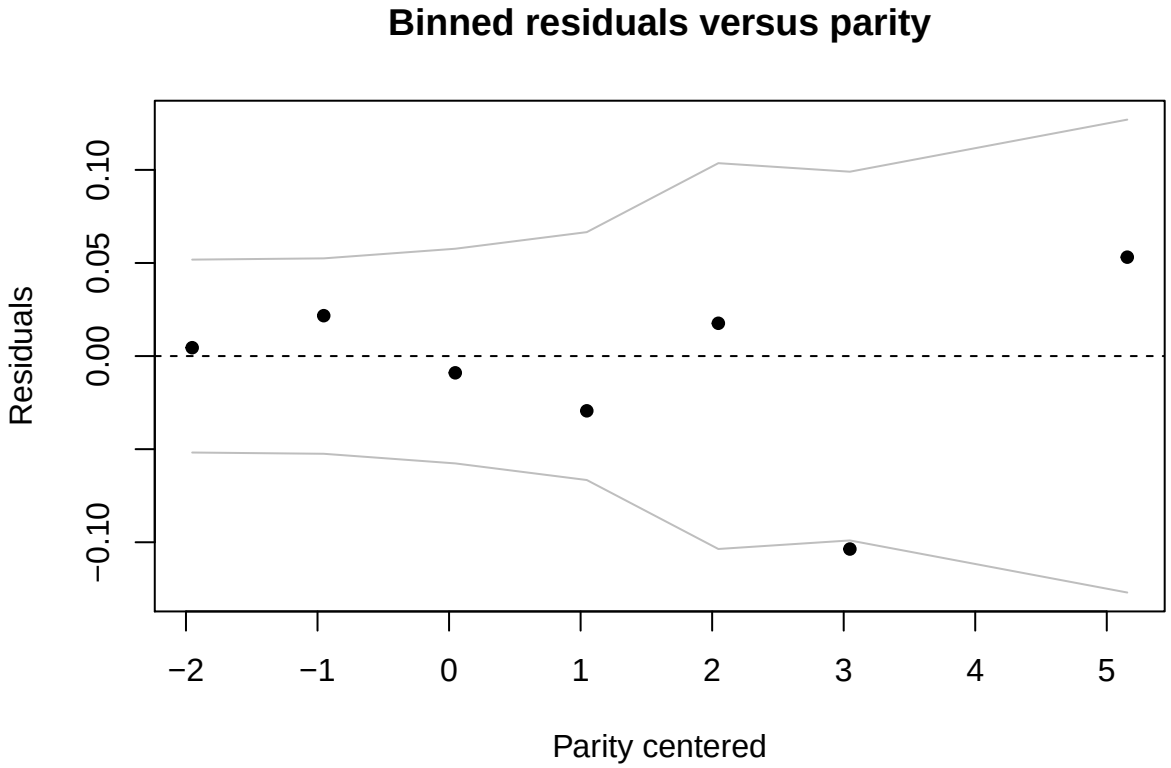
```r
maternal_data$parity.c <- maternal_data$parity - mean(maternal_data$parity)
maternal_data$mage.c <- maternal_data$mage - mean(maternal_data$mage)
maternal_data$mht.c <- maternal_data$mht - mean(maternal_data$mht)
maternal_data$mpregwt.c <- maternal_data$mpregwt - mean(maternal_data$mpregwt)
maternal_data$inc.c <- maternal_data$inc - mean(maternal_data$inc)

logis_mat_1 <- glm(premature~parity.c + mage.c + mht.c + mpregwt.c + inc.c + who_smoke + as.factor(mrac
summary(logis_mat_1)
```

```
##
## Call:
## glm(formula = premature ~ parity.c + mage.c + mht.c + mpregwt.c +
##     inc.c + who_smoke + as.factor(mrace_new) + as.factor(med),
##     family = binomial, data = maternal_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7422  -0.6743  -0.5574  -0.4066   2.4474
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.015552   0.963012  -1.055 0.291627
## parity.c               -0.018855   0.059855  -0.315 0.752748
## mage.c                  0.014886   0.020536   0.725 0.468528
## mht.c                  -0.026938   0.042223  -0.638 0.523481
## mpregwt.c              -0.011236   0.005525  -2.034 0.041985 *
## inc.c                   0.018709   0.043008   0.435 0.663559
## who_smoke               0.298560   0.185141   1.613 0.106829
## as.factor(mrace_new)6   0.131123   0.524208   0.250 0.802482
## as.factor(mrace_new)7   0.777520   0.232554   3.343 0.000828 ***
## as.factor(mrace_new)8   0.829257   0.414786   1.999 0.045582 *
## as.factor(mrace_new)9  -0.764453   1.054252  -0.725 0.468382
## as.factor(med)1        -0.356227   0.975016  -0.365 0.714846
## as.factor(med)2        -0.755554   0.962676  -0.785 0.432543
## as.factor(med)3        -0.621912   1.009737  -0.616 0.537951
## as.factor(med)4        -1.398608   0.979473  -1.428 0.153315
## as.factor(med)5        -0.971331   0.980907  -0.990 0.322058
## as.factor(med)7         1.953419   1.492183   1.309 0.190500
## ---
## 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: 794.50  on 852  degrees of freedom
## AIC: 828.5
##
## Number of Fisher Scoring iterations: 5
```

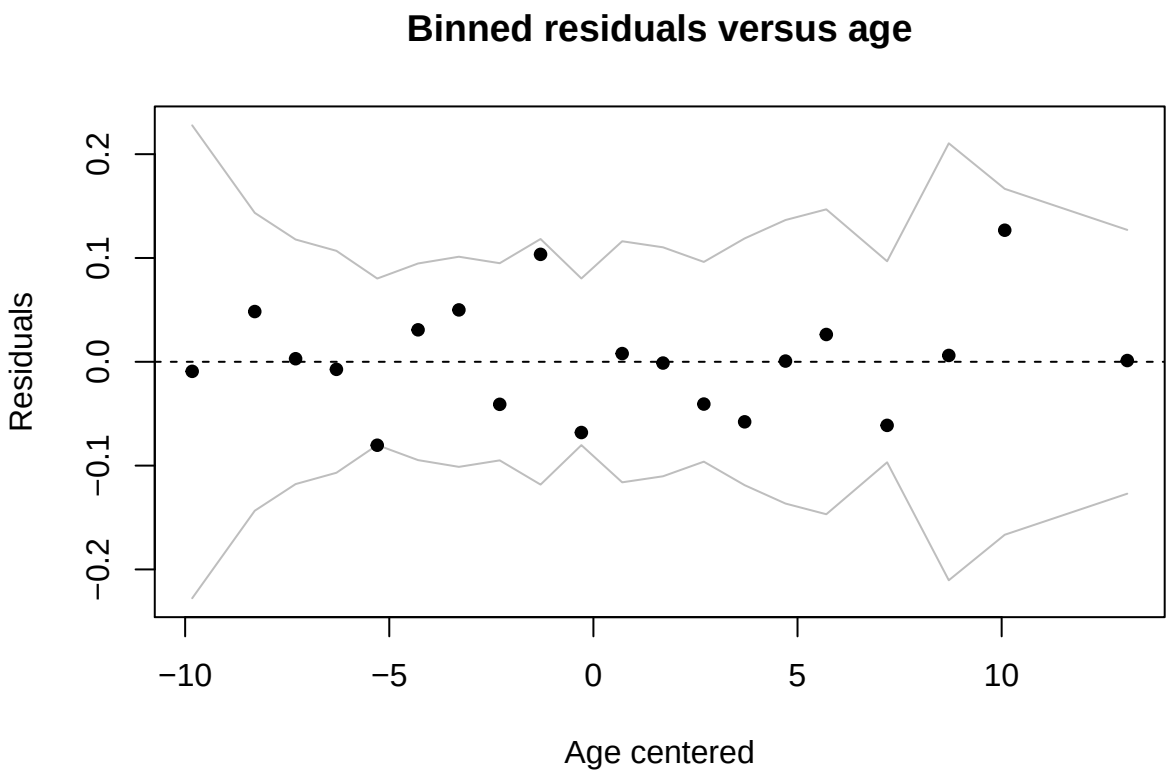
model diagnostics

We first do binned residual plots for numeric variables: We don't find any noticeable patterns.

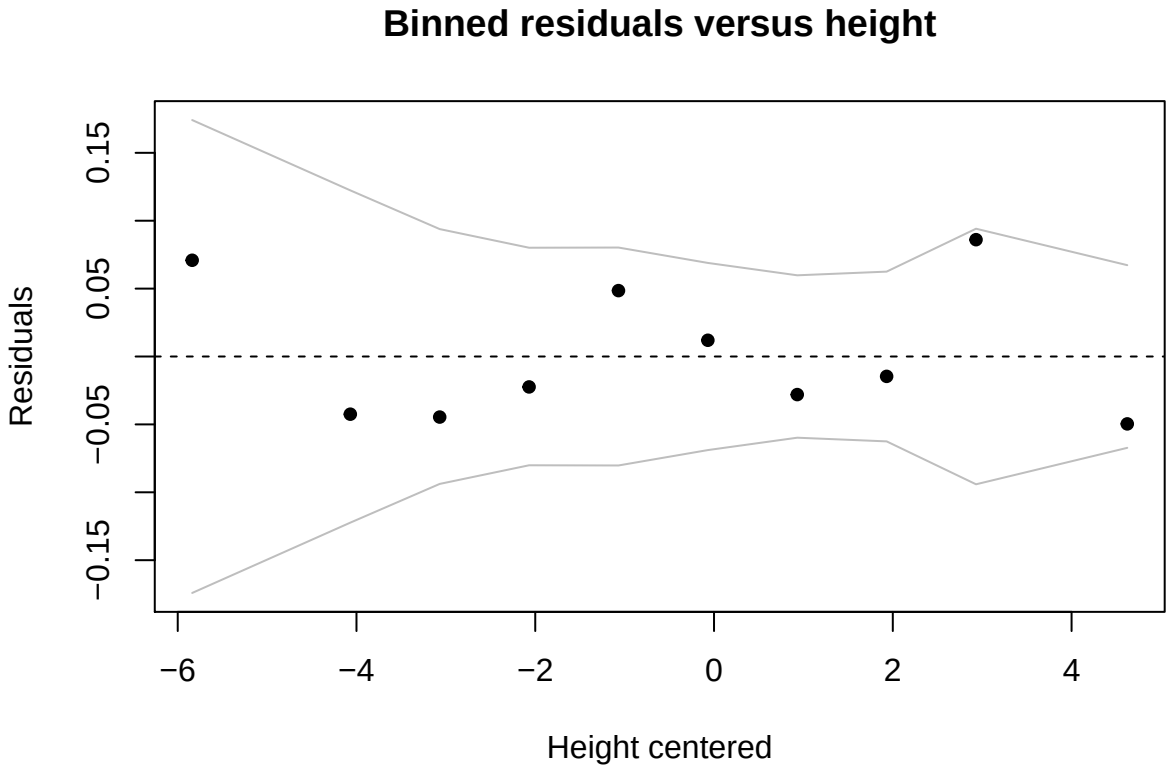
```
rawresid1 = maternal_data$premature - fitted(logis_mat_1)
binnedplot(x=maternal_data$parity.c, y = rawresid1, xlab = "Parity centered", ylab = "Residuals",
main = "Binned residuals versus parity")
```



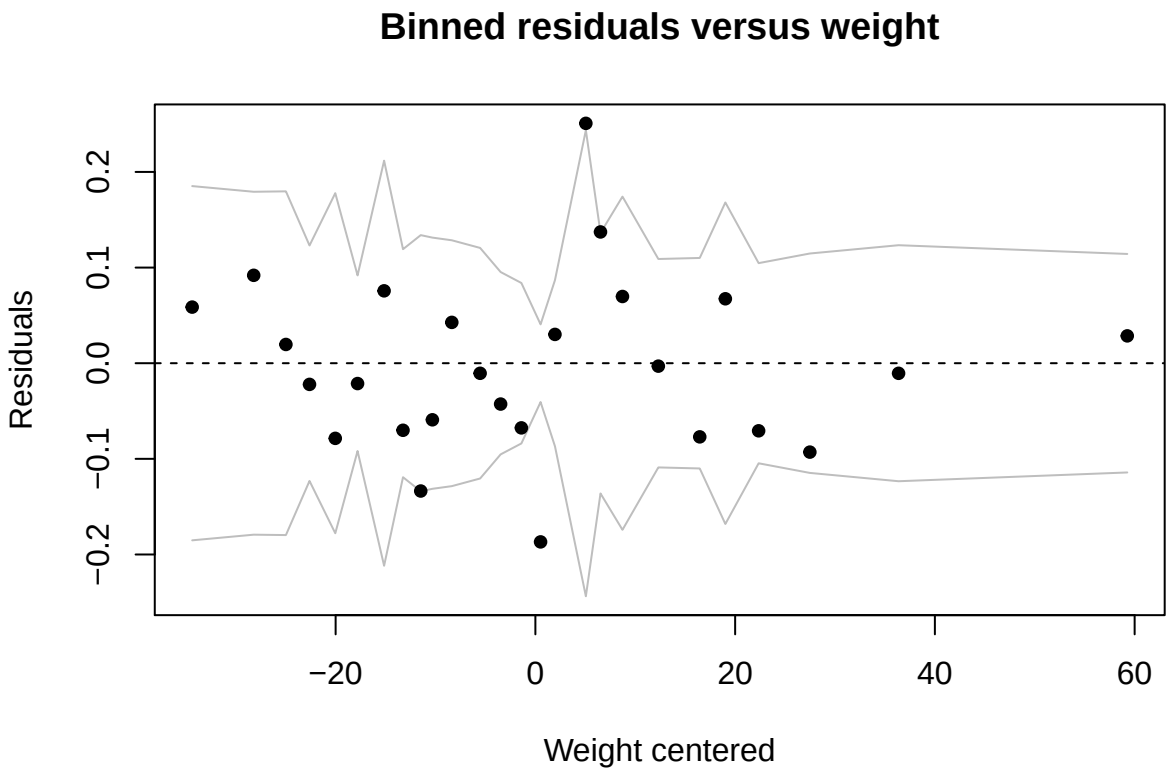

```
binnedplot(x=maternal_data$mage.c, y = rawresid1, xlab = "Age centered", ylab = "Residuals",
main = "Binned residuals versus age")
```



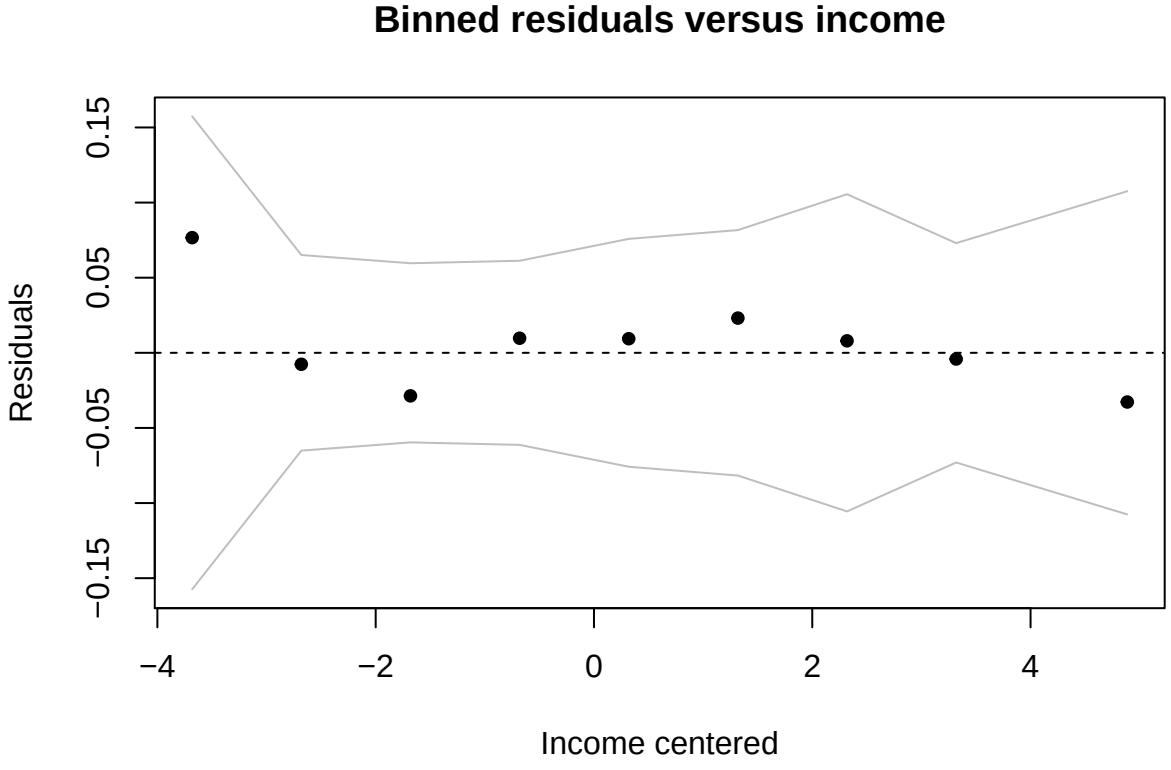

```
binnedplot(x=maternal_data$mht.c, y = rawresid1, xlab = "Height centered", ylab = "Residuals",
main = "Binned residuals versus height")
```




```
binnedplot(x=maternal_data$mpregwt.c, y = rawresid1, xlab = "Weight centered", ylab = "Residuals",
main = "Binned residuals versus weight")
```




```
binnedplot(x=maternal_data$inc.c, y = rawresid1, xlab = "Income centered", ylab = "Residuals",
main = "Binned residuals versus income")
```



Then look at average residuals by dummy variables using the tapply command: Nothing specific, except med level 7, which has few data, so we ignore that.

```
tapply(rawresid1, maternal_data$who_smoke, mean)
```

```
##             0             1
## -1.845932e-13 -6.418171e-14
```

```
tapply(rawresid1, maternal_data$mrace_new, mean)
```

```
##             5             6             7             8             9
##  1.325336e-17  6.533858e-16 -2.685290e-16 -1.649327e-16 -7.457275e-12
```

```
tapply(rawresid1, maternal_data$med, mean)
```

```
##             0             1             2             3             4
##  4.585221e-15 -1.682536e-13 -1.276143e-13 -1.629276e-13 -2.040208e-13
##             5             7
##  1.926044e-17 -8.326673e-17
```

Then do the confusion matrix with .3 threshold and .4 threshold, nothing specific:

```
threshold = 0.3
table(maternal_data$premature, logis_mat_1$fitted > threshold)
```

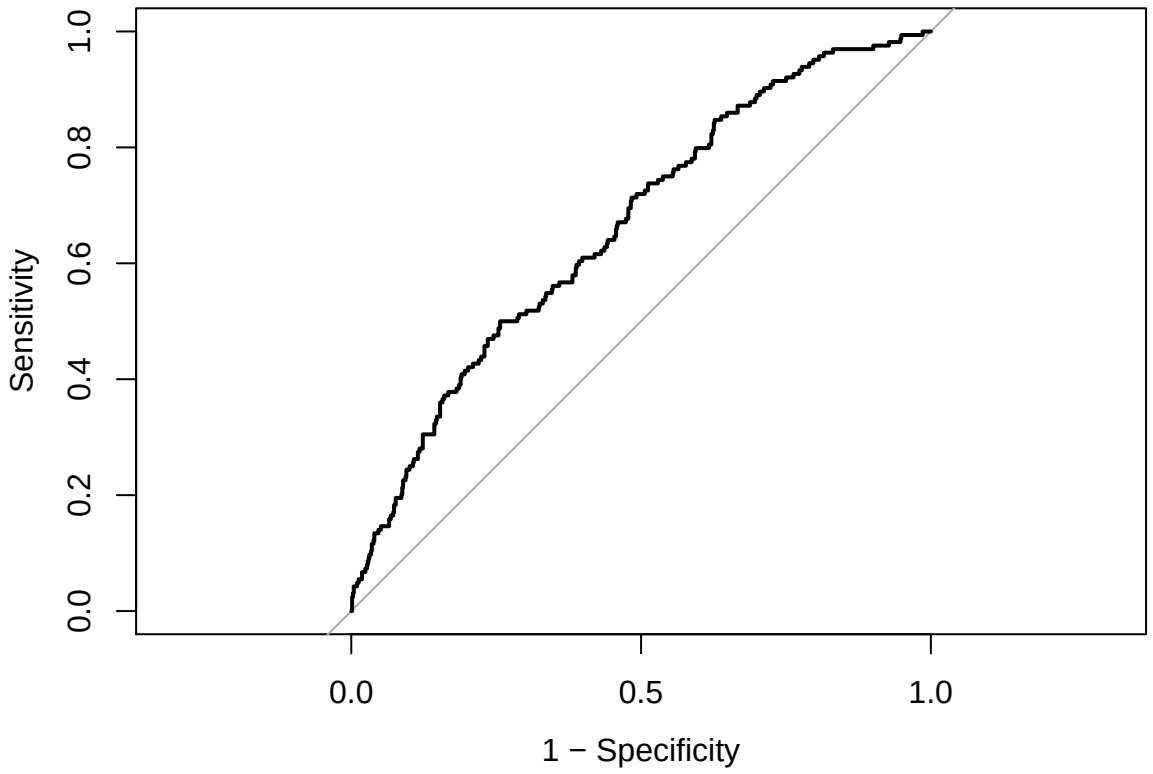
```
##
##     FALSE TRUE
##   0   638   67
##   1   125   39
```

```
threshold = 0.4
table(maternal_data$premature, logis_mat_1$fitted > threshold)
```

```
##
##     FALSE TRUE
##   0   689   16
##   1   153   11
```

Then look at ROC curve: We didn't find specific pattern from the model diagnostics, so we decide not to do transformations. We got area under the curve value: 0.6621.

```
roc(maternal_data$premature, fitted(logis_mat_1), plot=T, legacy.axes=T)
```




```
##
## Call:
## roc.default(response = maternal_data$premature, predictor = fitted(logis_mat_1),     plot = T, legacy
##
## Data: fitted(logis_mat_1) in 705 controls (maternal_data$premature 0) < 164 cases (maternal_data$prem
## Area under the curve: 0.6621
```

Then we look at if there are interactions between variables: For question 2, we need to first check if there is interaction between smoking and mother's race. We try the model with the interaction:

```
logis_mat_2 <- glm(premature~parity.c + mage.c + mht.c + mpregwt.c + inc.c + who_smoke * as.factor(mrace
summary(logis_mat_2)
```

```
##
## Call:
## glm(formula = premature ~ parity.c + mage.c + mht.c + mpregwt.c +
##     inc.c + who_smoke * as.factor(mrace_new) + as.factor(med),
##     family = binomial, data = maternal_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7570  -0.6802  -0.5512  -0.3958   2.4899
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.064396   0.981566  -1.084 0.278194
## parity.c                      -0.028393   0.060385  -0.470 0.638209
## mage.c                         0.015910   0.020652   0.770 0.441081
## mht.c                         -0.032685   0.042785  -0.764 0.444903
## mpregwt.c                     -0.011428   0.005489  -2.082 0.037353 *
```

```
## inc.c                              0.021160   0.043054   0.491 0.623094
## who_smoke                          0.417801   0.229134   1.823 0.068243 .
## as.factor(mrace_new)6              0.181878   0.630329   0.289 0.772930
## as.factor(mrace_new)7              1.086085   0.315813   3.439 0.000584 ***
## as.factor(mrace_new)8              0.740483   0.500649   1.479 0.139128
## as.factor(mrace_new)9            -13.533701 412.570072  -0.033 0.973831
## as.factor(med)1                   -0.367430   0.992500  -0.370 0.711229
## as.factor(med)2                   -0.782404   0.983973  -0.795 0.426528
## as.factor(med)3                   -0.667123   1.032635  -0.646 0.518254
## as.factor(med)4                   -1.429465   0.998650  -1.431 0.152316
## as.factor(med)5                   -0.981000   1.000036  -0.981 0.326610
## as.factor(med)7                    1.961329   1.515627   1.294 0.195641
## who_smoke:as.factor(mrace_new)6   -0.092326   1.127852  -0.082 0.934758
## who_smoke:as.factor(mrace_new)7   -0.601374   0.427002  -1.408 0.159023
## who_smoke:as.factor(mrace_new)8    0.323221   0.847853   0.381 0.703038
## who_smoke:as.factor(mrace_new)9   14.513284 412.572014   0.035 0.971938
## ---
## 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: 789.01  on 848  degrees of freedom
## AIC: 831.01
##
## Number of Fisher Scoring iterations: 14
```

We do change in deviance test to see if the interaction is useful. We get a p value of 0.2408, so we don't find interaction between smoking and race variable.

```
anova(logis_mat_1, logis_mat_2, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: premature ~ parity.c + mage.c + mht.c + mpregwt.c + inc.c + who_smoke +
##     as.factor(mrace_new) + as.factor(med)
## Model 2: premature ~ parity.c + mage.c + mht.c + mpregwt.c + inc.c + who_smoke *
##     as.factor(mrace_new) + as.factor(med)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       852     794.50
## 2       848     789.01  4   5.4883   0.2408
```

Scientifically, it is plausible to think that there might be interactions among weight variable and education, because mother with higher education level might focus more on their baby and willing to spend more part of their income to reduce pre-term birth, so I might try the interaction between income and education:

```
logis_mat_3 <- glm(premature~parity.c + mage.c + mht.c + mpregwt.c + inc.c * as.factor(med) + who_smoke
summary(logis_mat_3)
```

```
##
## Call:
## glm(formula = premature ~ parity.c + mage.c + mht.c + mpregwt.c +
##     inc.c * as.factor(med) + who_smoke + as.factor(mrace_new),
##     family = binomial, data = maternal_data)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.3101  -0.6588  -0.5474  -0.4083   2.4884
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             3.273e+01  1.146e+03   0.029  0.97722
## parity.c               -3.154e-02  6.190e-02  -0.509  0.61041
## mage.c                  2.156e-02  2.098e-02   1.028  0.30415
## mht.c                  -9.101e-03  4.343e-02  -0.210  0.83402
## mpregwt.c              -1.237e-02  5.618e-03  -2.202  0.02764 *
## inc.c                   2.850e+01  8.879e+02   0.032  0.97439
## as.factor(med)1        -3.429e+01  1.146e+03  -0.030  0.97613
## as.factor(med)2        -3.448e+01  1.146e+03  -0.030  0.97600
## as.factor(med)3        -3.453e+01  1.146e+03  -0.030  0.97597
## as.factor(med)4        -3.514e+01  1.146e+03  -0.031  0.97554
## as.factor(med)5        -3.477e+01  1.146e+03  -0.030  0.97580
## as.factor(med)7         8.345e+00  3.283e+03   0.003  0.99797
## who_smoke               2.836e-01  1.870e-01   1.516  0.12942
## as.factor(mrace_new)6 -2.403e-02  5.479e-01  -0.044  0.96501
## as.factor(mrace_new)7  7.511e-01  2.375e-01   3.162  0.00157 **
## as.factor(mrace_new)8  8.020e-01  4.183e-01   1.917  0.05524 .
## as.factor(mrace_new)9 -7.979e-01  1.057e+00  -0.755  0.45033
## inc.c:as.factor(med)1 -2.868e+01  8.879e+02  -0.032  0.97423
## inc.c:as.factor(med)2 -2.853e+01  8.879e+02  -0.032  0.97437
## inc.c:as.factor(med)3 -2.829e+01  8.879e+02  -0.032  0.97458
## inc.c:as.factor(med)4 -2.842e+01  8.879e+02  -0.032  0.97446
## inc.c:as.factor(med)5 -2.845e+01  8.879e+02  -0.032  0.97444
## inc.c:as.factor(med)7 -1.310e+01  1.451e+03  -0.009  0.99279
## ---
## 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: 781.84  on 846  degrees of freedom
## AIC: 827.84
##
## Number of Fisher Scoring iterations: 15
```

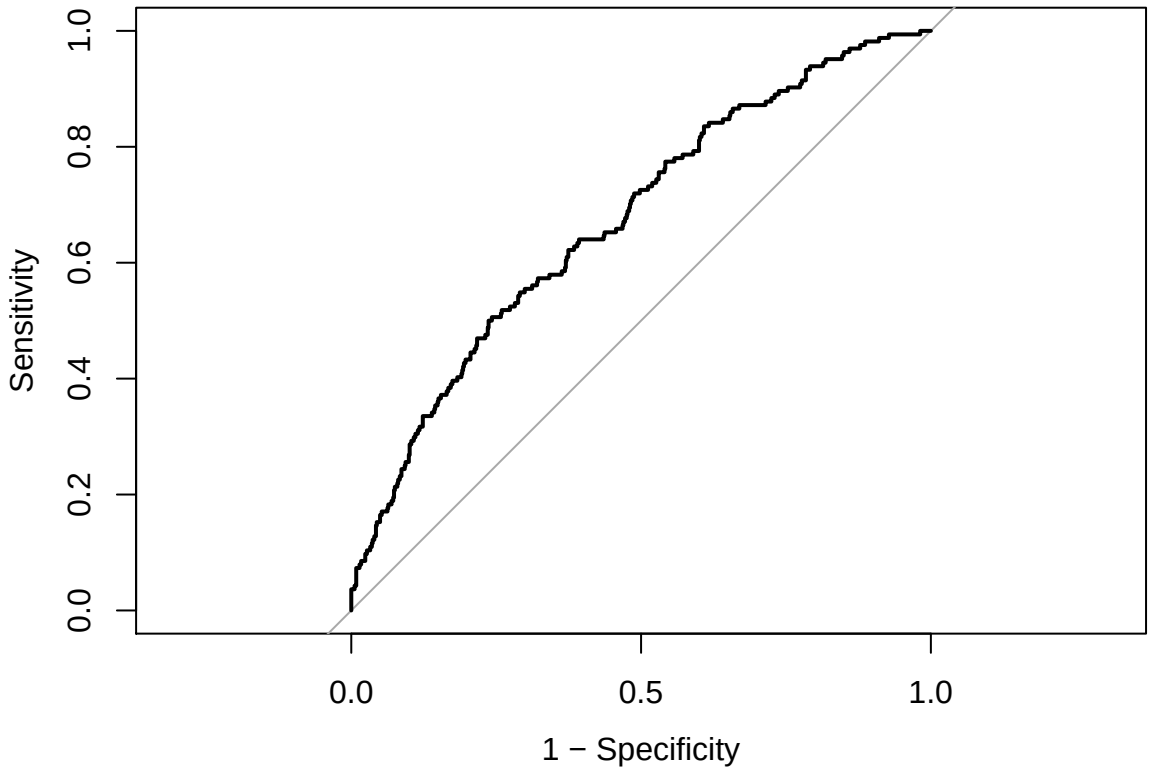
We do change in deviance test to see if the interaction is useful. We get a p value of 0.048, so we might find interaction between income and education.

```
anova(logis_mat_1, logis_mat_3, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: premature ~ parity.c + mage.c + mht.c + mpregwt.c + inc.c + who_smoke +
##     as.factor(mrace_new) + as.factor(med)
## Model 2: premature ~ parity.c + mage.c + mht.c + mpregwt.c + inc.c * as.factor(med) +
##     who_smoke + as.factor(mrace_new)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       852     794.50
## 2       846     781.84  6    12.66  0.04876 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We then check the ROC curve: We got a lightly better area under the curve value: 0.6729.

```
roc(maternal_data$premature, fitted(logis_mat_3), plot=T, legacy.axes=T)
```




```
##
## Call:
## roc.default(response = maternal_data$premature, predictor = fitted(logis_mat_3),     plot = T, legacy
##
## Data: fitted(logis_mat_3) in 705 controls (maternal_data$premature 0) < 164 cases (maternal_data$prem
## Area under the curve: 0.6729
```

Because the interaction we found is scientifically reasonable, and we got a way better ROC curve with that model, we decide to choose model_3 as our final model. Here is the model:

```
logis_mat_3 <- glm(premature~parity.c + mage.c + mht.c + mpregwt.c + inc.c * as.factor(med) + who_smoke
summary(logis_mat_3)
```

```
##
## Call:
## glm(formula = premature ~ parity.c + mage.c + mht.c + mpregwt.c +
##     inc.c * as.factor(med) + who_smoke + as.factor(mrace_new),
##     family = binomial, data = maternal_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3101  -0.6588  -0.5474  -0.4083   2.4884
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             3.273e+01  1.146e+03   0.029  0.97722
## parity.c               -3.154e-02  6.190e-02  -0.509  0.61041
```

```
## mage.c                 2.156e-02  2.098e-02   1.028  0.30415
## mht.c                 -9.101e-03  4.343e-02  -0.210  0.83402
## mpregwt.c             -1.237e-02  5.618e-03  -2.202  0.02764 *
## inc.c                  2.850e+01  8.879e+02   0.032  0.97439
## as.factor(med)1       -3.429e+01  1.146e+03  -0.030  0.97613
## as.factor(med)2       -3.448e+01  1.146e+03  -0.030  0.97600
## as.factor(med)3       -3.453e+01  1.146e+03  -0.030  0.97597
## as.factor(med)4       -3.514e+01  1.146e+03  -0.031  0.97554
## as.factor(med)5       -3.477e+01  1.146e+03  -0.030  0.97580
## as.factor(med)7        8.345e+00  3.283e+03   0.003  0.99797
## who_smoke              2.836e-01  1.870e-01   1.516  0.12942
## as.factor(mrace_new)6 -2.403e-02  5.479e-01  -0.044  0.96501
## as.factor(mrace_new)7  7.511e-01  2.375e-01   3.162  0.00157 **
## as.factor(mrace_new)8  8.020e-01  4.183e-01   1.917  0.05524 .
## as.factor(mrace_new)9 -7.979e-01  1.057e+00  -0.755  0.45033
## inc.c:as.factor(med)1 -2.868e+01  8.879e+02  -0.032  0.97423
## inc.c:as.factor(med)2 -2.853e+01  8.879e+02  -0.032  0.97437
## inc.c:as.factor(med)3 -2.829e+01  8.879e+02  -0.032  0.97458
## inc.c:as.factor(med)4 -2.842e+01  8.879e+02  -0.032  0.97446
## inc.c:as.factor(med)5 -2.845e+01  8.879e+02  -0.032  0.97444
## inc.c:as.factor(med)7 -1.310e+01  1.451e+03  -0.009  0.99279
## ---
## 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: 781.84  on 846  degrees of freedom
## AIC: 827.84
##
## Number of Fisher Scoring iterations: 15
```

And we take exponential for the coefficients and confidence intervals to get the odds:

```
exp(logis_mat_3$coefficients)
```

```
##           (Intercept)              parity.c                 mage.c
##          1.642861e+14          9.689557e-01          1.021791e+00
##                 mht.c               mpregwt.c                 inc.c
##          9.909404e-01          9.877046e-01          2.392367e+12
##       as.factor(med)1       as.factor(med)2       as.factor(med)3
##          1.282398e-15          1.061449e-15          1.012166e-15
##       as.factor(med)4       as.factor(med)5       as.factor(med)7
##          5.485082e-16          7.951732e-16          4.209359e+03
##             who_smoke as.factor(mrace_new)6 as.factor(mrace_new)7
##          1.327880e+00          9.762533e-01          2.119283e+00
## as.factor(mrace_new)8 as.factor(mrace_new)9 inc.c:as.factor(med)1
##          2.229899e+00          4.502620e-01          3.510118e-13
## inc.c:as.factor(med)2 inc.c:as.factor(med)3 inc.c:as.factor(med)4
##          4.085311e-13          5.152748e-13          4.535009e-13
## inc.c:as.factor(med)5 inc.c:as.factor(med)7
##          4.418542e-13          2.043168e-06
```

```
exp(confint(logis_mat_3))
```

```
##                                2.5 %        97.5 %
```

```
## (Intercept)           5.877729e+215  7.092113e+55
## parity.c               8.569334e-01  1.092963e+00
## mage.c                 9.800725e-01  1.064270e+00
## mht.c                  9.103694e-01  1.079630e+00
## mpregwt.c              9.765752e-01  9.983459e-01
## inc.c                 1.836551e+212 3.803876e+174
## as.factor(med)1        1.654980e-89 1.610502e-197
## as.factor(med)2        1.564532e-45 1.969778e-196
## as.factor(med)3        1.551941e-86 6.347622e-119
## as.factor(med)4        1.851500e-43 2.676122e-197
## as.factor(med)5        3.207468e-90 3.272757e-201
## as.factor(med)7        2.486090e-08  4.928507e+14
## who_smoke              9.206806e-01  1.918340e+00
## as.factor(mrace_new)6  2.978351e-01  2.659879e+00
## as.factor(mrace_new)7  1.325023e+00  3.368739e+00
## as.factor(mrace_new)8  9.545518e-01  4.984606e+00
## as.factor(mrace_new)9  2.432247e-02  2.391258e+00
## inc.c:as.factor(med)1 2.986412e-172 9.083281e-152
## inc.c:as.factor(med)2 5.385344e-176 2.762417e-226
## inc.c:as.factor(med)3 5.640739e-170 2.634480e-110
## inc.c:as.factor(med)4 1.092729e-168  4.458507e-84
## inc.c:as.factor(med)5 4.392904e-203 1.078450e-259
## inc.c:as.factor(med)7  1.628644e-11  1.533865e-01
```

Interpretation:

Answer for question 1:

Holding other variables constant, mothers who smoke tend to have 32.79% higher odds of having pre-term birth than mothers who do not smoke. The 95% confidence interval for the odds of pre-term birth for smokering mothers are from 7.93% lower to 91.83% higher (92.07%, 191.83%) than non-smoking mothers.

Answer for question 2:

We did change in deviance test to see if the interaction between smoking and race is valid. We got a p value of $0.2408 > 0.05$, so we don't find interaction between smoking and race variable.

Answer for question 3:

I found several interesting associations with the odds of pre-term birth:

1, We found it interesting that holding other variables constant, when pre-pregnancy weight increase by 1 pound, the odds of getting a pre-term birth in fact reduce by 1.23%, with a confidence interval of (0.17%, 2.34%) reducing odds of having pre-term birth.

2, We found another thing that holding other variables constant, mothers of race level 7 and 8, which is black and asian, in fact have way much higher odds of pre-term birth than white mothers. Black mothers tend to have 111.93% higher odds (more than 2 times of the odds) of having a pre-term birth than white mothers; asian mothers tend to have 122.99% higher odds (more than 2 times of the odds) of having a pre-term birth than white mothers.